Anàlisi estadística composicional en l'avaluació dels serveis d'Aigua, Sanejament i Higiene en països d'ingressos baixos

TFM

Màster en Tecnologia per al
Desenvolupament Humà i la Cooperació

Enric Lloret i Bosch
Directora: Maribel Ortego Martínez
Codirector: Agustí Pérez-Foguet

Universitat Politècnica de Catalunya (UPC)
Setembre 2016

Als meus pares que em crearen i em criaren lliure i crític. Per ser i estar, sempre, incondicionalment.

"Si tens un fill, ensenya'l a ser lliure. Encara que siga a costa teua.
En realitat, haurà de ser a costa teua"
J.F.

A les que han estat en algun moment: no som res més que assemblatges d'altres, de vivències, de pensaments, de llocs, de coses.

"Los científicos dicen que estamos hechos de átomos pero a mí
un pajarito me contó que estamos hechos de historias"
E.G.

A la gent de Madrid i Manhiça, que em donaren l'oportunitat de llevar-me una espineta clavada i me'n clavaren quatre mil més. Sense elles, aquesta Barcelona no existiria.

«Al fin y al cabo, somos lo que hacemos para cambiar lo que somos»
E.G.

A la gent del màster, companyes d'estudi i de vida. A Agustí, Ricard i Jordi que en algun moment els he marejat amb les meues històries. A Maribel, per tant, per tot.

« No faces de la teua ignorància un argument »

J.F.

A les meues de Sueca, que són la base, el fonament, l'essència.

A Lina, per fer-ho tant fàcil sempre tot, per estar sempre disposta a crear junts, per creure, per fer equip, per ixe somriure, pel que hi ha dins i fora, pel que vindrà, per ser com eres.

INDEX

1 RESUM / ABSTRACT

El Programa conjunt de monitoratge per a abastament d'aigua i sanejament (JMP per les seues sigles angleses) de l'Organització Mundial de la Salut, com a organisme encarregat del monitoratge del progrés per a la consecució de la meta 7c dels Objectius Del Mil·lenni (ODM), ha estat realitzant estimacions de cobertura pel que fa a l'accés a un font d'aigua de beure i a les instal·lacions de sanejament a escala país, regional i mundial. Per fer-ho ha utilitzat una metodologia basada en el model de regressió lineal simple aplicat directament sobre les dades observades, amb una lleugera modificació per evitar la predicció de valors impossibles (proporcions negatives o superiors al 100%) en escenaris futurs o passats. Arribats a l'escenari final dels ODM, l'any 2015, s'ha plantejat una revisió d'aquesta metodologia. Diversos equips investigadors han proposat altres models d'ajust directament sobre les dades observades, sense tenir en compte que ni aquests models ni el model del JMP respecten la natura composicional de les dades, donant lloc a prediccions que podrien ser errònies. A aquesta tesina analitzem la metodologia composicional aplicada a les dades del JMP i particularitzada a un conjunt de països de l'Àfrica subsahariana, desenvolupant de manera completa tot el procediment per a les dades d'accés a l'aigua a Ghana en entorn urbà i rural. Els valors nuls a alguna de les parts de la composició necessiten ser tractats prèviament. Es mostren dos exemples d'aquest tractament i la influència dels valors de substitució adoptats en les prediccions del model. La metodologia composicional proposada es tracta d'una metodologia senzilla, sistematitzable (una vegada "eliminats" els zeros de la sèrie original) i que permet realitzar prediccions a futur conservant la natura composicional de les dades, pel que res impedeix que siga considerada com a una alternativa sòlida al mètode del JMP.

ABSTRACT

The WHO/UNICEF Joint Monitoring Program (JMP) is the institution officially assigned to monitor progress to reach MDG target 7c. JMP has been providing coverage estimations on drinking water access and sanitation facilities access since 2000. JMP predictions are based on a linear regression method directly applied to the available data set, slightly modified in order to avoid impossible predictions such as negative (or above 100%) percentages in past or future scenarios. 2015-MDG period has been reached, JMP has launched a revision of the method: some research groups have suggested some alternatives that are applied (again) directly to the original set of data, without taking into account the compositional nature of that data. Consequently, as pointed out by many authors, the results given might be wrong. In this TFM, a Compositional Data (CoDa) approach is analyzed as an alternative to the current JMP method. It is applied to JMP data set, particularly for some South African region countries, and its results analyzed. As an example, the whole method is developed step by step for a country (water data access in Ghana, in both rural and urban context). The presence of zero values in the original data set needs a special treatment before performing a CoDa analyses. Two examples of zero treatment and the influence of the zero replacement value are shown.

In conclusion, CoDa approach is a simple and easy to be systematized (after replacing the zero values). It allows us to obtain predictions on future scenarios preserving the CoDa nature of data: nothing prevents it from being considered as a solid alternative to JMP method.

2 DEFINICIÓ D'OBJECTIUS

Els objectius d'aquest treball són els següents:

1. Comprendre la metodologia utilitzada pel JMP (*Joint Monitoring Program*) per estimar la situació en relació a l'accés a l'aigua i al sanejament i avaluar el grau de compliment de la meta 7c dels Objectius del Mil·lenni (ODM).

2. Analitzar la problemàtica del tractament de dades de tipus composicional amb els mètodes estadístics clàssics i, en conseqüència, la importància de l'anàlisi composicional en aquests casos.

3. Mostrar i analitzar de forma crítica les diferències metodològiques entre el model lineal (model de base del JMP) i el derivat de l'anàlisi composicional, mitjançant la seua aplicació a les sèries de dades disponibles per a distints països de l'Àfrica Sub-Sahariana.

4. Analitzar la importància de les dades d'entrada (resultat del treball de recollida de dades realitzat) i específicament la influència de la presència de zeros a la sèrie. Interpretació d'aquests zeros i bases per al seu tractament.

5. Identificar punts forts i dèbils en la consideració de transformacions per a dades composicionals en la obtenció d'estimacions per aigua i sanejament.

3 ANTECEDENTS

D'acord amb les dades de l'Organització Mundial de la Salut (OMS o WHO per les seves sigles en anglès) corresponents a 2015, dos mil cinc-cents milions de persones no tenien accés a un sanejament millorat i mil milions no tenien accés a cap tipus d'instal·lació. Aquests mil milions de persones sense cap instal·lació de sanejament continuaven defecant en fosses, darrere d'arbustos o en masses d'aigua superficial, sense dignitat ni privacitat. Nou de cada deu persones que defecaven a l'aire lliure vivien en zones rurals, però el nombre d'aquests en zones urbanes anava en augment. A més a més, 748 milions de persones - majoritàriament els pobres i marginals- encara no tenien accés a una font millorada d'aigua per beure. D'aquests, gairebé un quart (173 milions) s'abastien d'aigua superficial no tractada, i més del 90% vivien en zona rural (WHO/UNICEF, 2015).

Per contribuir a acabar amb aquesta situació, l'aigua i el sanejament foren específicament inclosos dins dels objectius de desenvolupament del mil·lenni (ODM), formulats a l'any 2000. Dels vuit objectius establerts, el setè d'ells és "Garantir la sostenibilitat del Medi Ambient" i s'estructura en quatre Metes. D'aquestes, la meta 7c estableix la necessitat de reduir a la meitat, al 2015 i respecte a les dades de 1990, la proporció de persones sense accés a una font millorada d'aigua potable i a una instal·lació bàsica de sanejament. Aquest objectiu, acceptat a nivell internacional, ha suposat un fort vector de desenvolupament: tant governs nacionals com agències donants han centrat l'atenció en el progrés per aconseguir-lo i en els distints objectius relacionats amb aquest (WHO/UNICEF, 2015).

El Programa Conjunt de Monitoratge per a Abastament d'Aigua i Sanejament (JMP per les seues sigles angleses) de l'OMS/UNICEF, ha estat publicant estimacions sobre taxes de cobertura a nivell global i tendències des del 1990, i és l'organisme oficialment designat per tal de monitorar el progrés cap a l'ODM 7c per a aigua i sanejament. Per tal de fer-ho, el JMP no tan sols publica dades sobre els últims sondejos realitzats sinó que, a més, realitza estimacions a futur basades en el model de regressió lineal simple. La regressió lineal permet tractar amb dades amb petites diferències de cobertura, proporciona estimacions de dades per a anys on no es disposa d'aquestes, i és relativament fàcil d'explicar per a agents polítics i professionals de les entitats encarregades de la prestació de serveis d'aigua i sanejament.

Davant la fi del període dels ODM en 2015, al si del JMP s'impulsà a les acaballes de 2014 una discussió internacional (a través d'una reunió d'experts celebrada a Nova York) al voltant de l'agenda post 2015, objectius a assolir i indicadors, així com una revisió del mètode actual utilitzat pel JMP per obtenir les estimacions de cobertura (WHO & UNICEF, 2014). La UPC, a través de l'Institut de Sostenibilitat, estigué present a aquesta trobada.

Cal tenir en compte que des de l'inici del monitoratge de l'estat de la situació en relació a l'aigua i el sanejament a escala internacional pel sistema de Nacions Unides, els mètodes utilitzats han evolucionat de manera considerable i el volum i la quantitat de dades disponibles ha crescut ràpidament. A pesar de que les dades recollides pel JMP (i les estimacions derivades a partir de l'explotació d'aquestes) han sigut de gran ajuda per als distints actors involucrats, distints estudis han assenyalat limitacions en el mètode actual, així com possibilitats de millora d'aquest (Bartram et al., 2014)

En aquest context, diversos grups investigadors a nivell internacional han estat comparant les estimacions de cobertura obtingudes a partir de la metodologia utilitzada pel JMP amb estimacions obtingudes a partir de metodologies alternatives a aquesta. Entre aquestes metodologies alternatives s'hi troba l'anàlisi de les dades des d'una aproximació distinta a l'estadística clàssica, donat que el tipus de dades disponibles (dades composicionals) presenten unes particularitats específiques que condicionen la seua anàlisi estadística.

El tema central d'aquesta tesina és precisament aquest: l'anàlisi de les dades d'acord amb la seua natura composicional i la comparació amb el mètode utilitzat pel JMP dels resultats que se'n deriven.

Així, en primer lloc (apartat 4) es descriurà de forma breu el mètode utilitzat pel JMP per a la realització de les estimacions a futur a partir de les dades disponibles.
Una vegada descrit aquest, enumerarem (apartat 5) altres metodologies que han estat proposades com a alternatives al mètode actual del JMP. Entre aquestes introduirem de manera breu l'anàlisi composicional, que desenvoluparem més endavant a l'apartat 7.
A l'apartat sisè (6) es descriuran les sèries de dades d'aigua i sanejament facilitades pel JMP i que, després de ser reorganitzades, suposen les dades de partida per a l'anàlisi que hem realitzat.
Seguidament, a l'apartat 7, es realitzarà una descripció metodològica del mètode actual del JMP i de l'anàlisi composicional per a posteriorment, a l'apartat 8, desenvolupar aquesta metodologia sobre les dades disponibles i analitzar els resultats obtinguts.
A l'apartat 9 es suggeriran les línies d'investigació que donarien continuïtat a aquesta tesina i que permetrien aprofundir en el tema; i finalment, a l'apartat 10, es realitzaran les conclusions d'aquest treball.

4 EL MÈTODE DEL JMP

4.1 Definicions

El JMP proveeix estimacions de cobertura, comparables entre països i al llarg del temps, i monitora els progressos per tal d'assolir la meta 7c dels Objectius del Mil·lenni sobre aigua i sanejament. Per a l'elaboració d'informes d'estat d'aquest procés, el JMP ha adoptat les següents definicions (WHO/UNICEF, 2015):

- Una font millorada d'aigua potable és aquella que, per la natura de la seua construcció, està protegida adequadament de la contaminació exterior, particularment de contaminació d'origen fecal.
- Una instal·lació de sanejament millorada és aquella que impedeix de manera higiènica el contacte de les persones amb l'excreta humana. Les instal·lacions de sanejament compartides amb altres llars no són considerades com millorades.

El JMP ha establert un conjunt estàndard de categories per a l'anàlisi de dades a nivell nacional, dades sobre les quals es basen les tendències i estimacions d'indicadors en relació als objectius del mil·lenni. Aquestes categories s'organitzen en distints nivells, constituint el que anomenem com escales d'aigua i de sanejament. Aquestes escales d'aigua (figura 1) i sanejament (figura 2) no únicament distingeixen entre instal·lacions millorades i no millorades sinó que també proveeixen informació addicional sobre els nivells de servei.

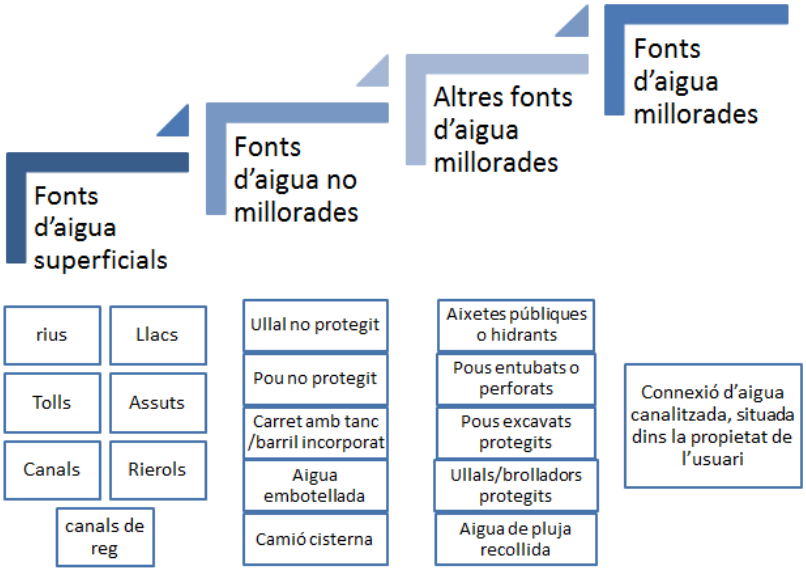


Figura 1. Escala d'aigua. Font: elaboració pròpia a partir de WHO/UNICEF, 2015

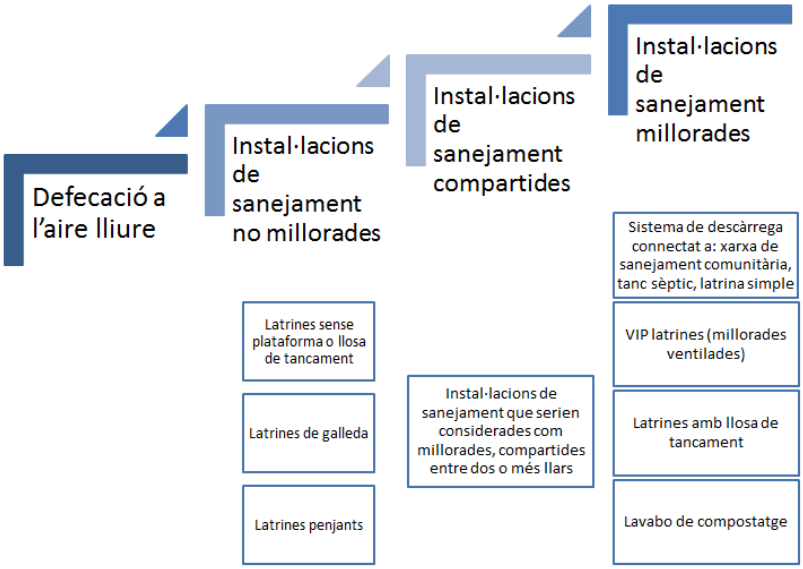


Figura 2. Escala de sanejament. Font: elaboració pròpia a partir de WHO/UNICEF, 2015

4.2 Origen de les dades

Les dades del JMP provenen de tres fonts d'informació principals : enquestes domiciliàries (per llars), censos i informes administratius sobre l'ús de fonts d'aigua potable i d'instal·lacions de sanejament, i dades demogràfiques de la Divisió de Població de les Nacions Unides (WHO/UNICEF, 2015).

Cal destacar la gran heterogeneïtat pel que fa al nombre de dades disponibles entre països. Així, per a un indicador determinat, alguns països disposen de poques dades i altres, de més de 20 (Fuller, Goldstick, Bartram, & Eisenberg, 2016).

Els principals sondejos en els quals es basa el JMP són els següents (WHO/UNICEF, n.d.):

a. Enquestes demogràfiques i de Salut (*Demographic and Health Surveys -DHS-*):
Es tracta d'enquestes a nivell domiciliar, realitzades per a cada país i finançades per l'Agència dels Estats Units per al Desenvolupament Internacional (USAID), que proporcionen dades en relació a un ampli ventall d'indicadors d'avaluació d'impacte i de monitoratge, en els sectors de població (demografia), salut i nutrició. El tamany de les mostres està entre 2 000 i 30 000 llars, i els sondejos es realitzen en més de 75 països, cada 5 anys aproximadament.

b. Enquestes d'Indicadors Múltiples per Conglomerats (*Multiple Indicator Cluster Surveys (MICS)*)
UNICEF dóna assistència als països pel que fa a la presa de dades i la seua anàlisi per completar les sèries de dades per al monitoratge de la situació dels infants i les dones, a través de la iniciativa internacional d'enquestes a escala domèstica d'Indicadors Múltiples per Conglomerats (MICS). Des de mitjans dels anys 1990, el MICS ha permès a molts països produir estimacions estadístiques i internacionalment comparables d'una sèrie d'indicadors en les àrees de la Salut (incloent Aigua, Sanejament i Higiene), l'educació, la protecció als infants i la Sida.

c. Enquestes de Salut Mundial (World Health Surveys (WHS))
L'Organització Mundial de la Salut ha desenvolupat i implementat un programa de mostreig i un mostreig mundial en matèria de salut, per tal de compilar informació sobre la salut de la

població i sobre com la inversió en sistemes de salut repercuteix a la millora d'aquesta; evidències sobre com els sistemes de salut estan funcionant en l'actualitat; i capacitar per al monitoratge de les dades d'entrada, l'evolució i els progressos aconseguits.

d. Enquestes de mesura dels estàndards de vida (*Living Standards Measurement Surveys (LSMS)*)
Es tracta d'una investigació que s'està duent a terme al Banc Mundial, generant dades a escala domiciliar.

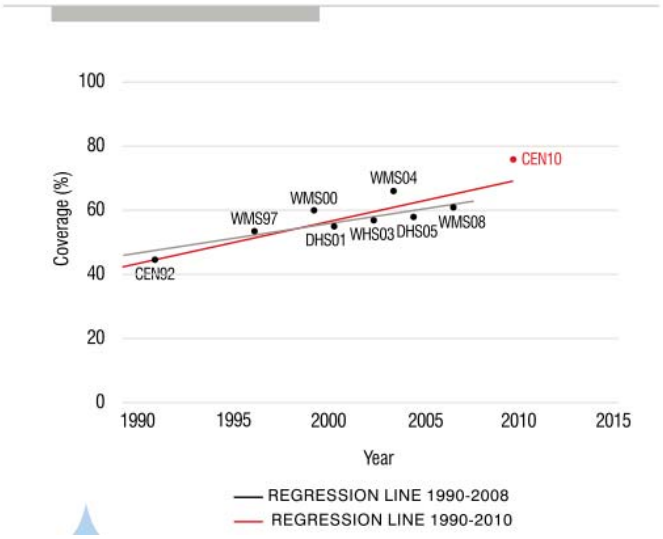
I pel que fa als censos:

- Censos de població i a escala domèstica (*Population and housing censuses*)
Les dades pel que fa a l'accés a l'aigua i al sanejament són normalment preses en censos a escala domèstica en la majoria de països en desenvolupament. Aquests censos són per tant una font de dades important per a les estimacions realitzades pel JMP. Les estimacions actuals es realitzen a partir de les dades recopilades en més de 250 censos.

4.3 Mètode d'estimació

A partir del conjunt de sondejos i censos disponibles, el JMP extreu les dades corresponents a deu variables. Es tracta de proporcions de llars rurals i urbanes que utilitzen: aigua canalitzada a dins de la propietat, aigua provinent de qualsevol font classificada com millorada (incloent-hi la canalitzada dins la propietat), aigua recollida d'una font superficial, qualsevol tipus d'instal·lació de sanejament classificada com millorada, i no utilització de cap instal·lació de sanejament (llars considerades com practicant la defecació a l'aire lliure) (Bartram et al., 2014).

Amb caràcter anual, a partir de les dades disponibles i per a controlar l'evolució de cada país en relació als ODM, el JMP realitza estimacions de cobertura per a cada país, basades en ajustar una recta de regressió a les sèries de dades de cadascuna de les deu variables descrites anteriorment. Un exemple d'aquest ajust per a un país determinat pot observar-se al gràfic 1.



Gràfic 1. Exemple d'ajust lineal per a un país. Font: WHO/UNICEF, 2015

El mètode utilitzat pel JMP per tal d'estimar els percentatges de gent que utilitzen un determinat tipus de font d'aigua per abastar-se o un d'instal·lació de sanejament és la regressió lineal simple (RLS) (WHO/UNICEF, 2015).

Les rectes de regressió s'extrapolen fins com a màxim dos anys abans del cens o sondeig més antic i dos anys després del més recent (sempre forçant que el resultat estiga comprés entre 0 i 100%) (Bartram et al., 2014). Més enllà d'aquests punts, les estimacions romanen invariables (pendent nul.la) durant com a molt 4 anys, excepte quan el percentatge es situa per baix del 5% o per dalt del 95%, casos on la línia s'estén de manera indefinida (WHO/UNICEF, 2015). Més enllà dels període de quatre anys abans i després de l'última dada disponible, les estimacions es recullen com a "no disponible".

Un exemple del tipus d'estimació realitzat d'acord amb el mètode del JMP, obtingut de Bartram et al., 2014 és el de la figura 3 on s'inclouen sis gràfics representatius de sis simulacions corresponents a un país qualsevol. Es disposa de dades d'una variable corresponents a un període de temps determinat i vol estimar-se el valor d'aquesta variable per als anys 1992 i 2012.

Al cas de la figura 3A, les dades disponibles són 9 punts corresponents al període 1992-2010. La recta de regressió ajustada per a les dades disponibles s'extrapola 2 anys abans de la primera dada (1992) i de l'última (2010), per obtenir les estimacions als anys 1990 (26.1%) i 2012 (57.6%).

Al cas de la figura 3B, es disposa també de 9 punts però ara durant període 1992-2008. L'estimació de la variable per al 1990 s'obté directament de l'extrapolació de la recta de regressió a aquest any (donat que correspon a 2 anys abans de la primera dada disponible) però per estimar el valor corresponent al 2012 es realitza una extrapolació de la recta 2 anys (fins el 2010) i dos anys més amb pendent plana. El valor estimat al 2012 seria de 58.1%.

Al cas de la figura 3C sols es disposa de 5 dades durant el període 1996-2006. S'extrapola la recta dos anys i seguidament tenim un període de 4 anys de pendent nul.la. Amb això, l'estimació per al 1990 és de 29.6% i per al 2012 de 54.3%.

Al cas 3D es disposa de 4 punts durant el període 1996-2004. La recta és extrapolada durant 2 anys i seguidament l'estimació dels 4 anys avant i enrere es realitzaria amb pendent nul.la. D'aquesta forma, l'estimació 1990 seria de 29% però l'estimació per al 2012 es consideraria com a no disponible ja que el 2012 està distanciat més de 6 anys del valor més proper.

El cas 3E representa el cas de pendent nul.la per percentatges inferiors al 5% i el 3F, el cas de pendent nul.la per percentatges superiors al 95%.

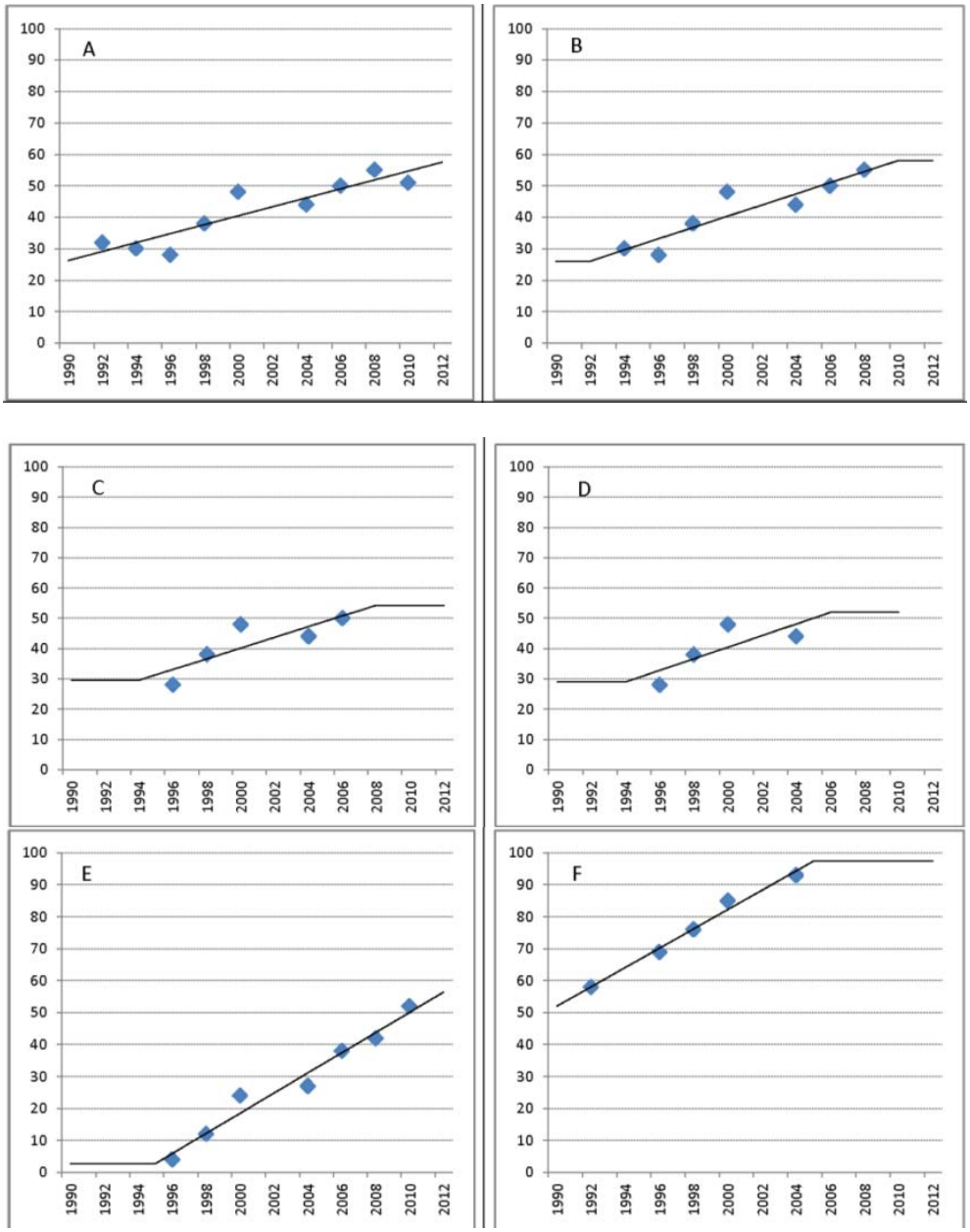


Figura 3. Simulació d'extrapolacions de la recta de regressió d'acord amb el mètode del JMP per a països amb dades no disponibles. Font : (Bartram et al., 2014)

Seguint el criteri descrit anteriorment, el JMP realitza estimacions de les següents variables:

- Dades d'aigua: Proporció de gent (% de cobertura) que utilitza les següents fonts:
  - Canalitzada a la propietat
  - Fonts millorades d'aigua (incloent la canalitzada a la propietat)
  - Fonts superficials
- Dades de Sanejament: s'estimen els percentatges de gent que:
  - Utilitza instal·lacions millorades (tant si són compartides com si no)
  - Defeca a l'aire lliure

La població restant utilitza fonts d'aigua no millorades i instal·lacions de sanejament no millorades, respectivament (WHO/UNICEF, 2015). Aquesta forma de fer garanteix que la proporció total siga el 100%.

Al mètode del JMP les regressions per a zones rurals i urbanes es consideren per separat, obtenint així valors desagregats per zones. Els valors corresponents al conjunt de la població de cada país s'estimen a partir dels valors desagregats anteriors (rural/urbà). Per fer-ho, es prenen coeficients de ponderació en funció de la relació de població vivint en zona urbana i rural respecte a la total de cada país, obtinguda a partir dels valors més recents de les estimacions de població realitzades per la Divisió de Població de les Nacions Unides (Bartram et al., 2014).

Els valors estimats de cadascuna de les variables considerades per al conjunt del país (població total, rural i urbana) són els que s'utilitzen per tal de realitzar les estimacions a escala nacional, regional i global.

En el cas particular de les dades de sanejament, el monitoratge de l'accés a una instal·lació de sanejament millorada presenta una particularitat. La definició adoptada pel JMP respecte aquest tipus d'instal·lacions és clara pel que fa a que l'ús no ha de ser compartit entre distintes llars. Tanmateix, les dades recollides pel JMP sobre accés a instal·lacions millorades es fixen únicament en la infraestructura i no en la seua utilització, és a dir, inclouen tant aquelles que són d'ús individualitzat com compartit. És per això que cal desagregar aquests valors.

Cal tenir en compte que les dades disponibles en relació al percentatge de llars que comparteixen instal·lacions de sanejament són minoritàries (i provinents de censos recents). El valor mitjà de les proporcions de població rural i urbana vivint en llars que comparteixen instal·lació de sanejament es calcula a partir dels sondejos on aquestes dades són disponibles. Aquests valors mitjans, expressats en percentatges, són les estimacions de cobertura d'ús compartit d'instal·lacions de sanejament, i es consideren constants en el temps (WHO/UNICEF, 2015) (Bartram et al., 2014).

5 MÈTODES ALTERNATIUS D'ESTIMACIÓ

El JMP ha estat criticat degut a la gran diferència entre les seues estimacions i les estimacions realitzades per les institucions nacionals dels propis països. Com assenyalen alguns autors (Bartram et al. 2014), aquestes diferències es deuen fonamentalment a dos factors: 1) el propi mètode del JMP, basat en un ajust lineal de la sèrie de dades; i 2) la diferència de criteris alhora de designar quin tipus de fonts o instal·lacions corresponen a cadascuna de les categories dins de l'escala d'aigua i sanejament (figures 1 i 2).

Les pròpies institucions responsables del JMP reconeixen els problemes associats al model de regressió lineal i justifiquen que l'increment del nombre de dades disponibles permet l'exploració de models més sofisticats (WHO/UNICEF, 2015).

En 2014, el propi JMP organitzà un taller d'experts per explorar mètodes alternatius a la RLS i la seua aplicació potencial per al període post-ODM (post 2015). Tal i com s'indicava a l'informe sorgit d'aquell taller (WHO & UNICEF, 2014), s'evidencià la presència de patrons no lineals en alguns països tot i que, per a la majoria d'ells falta tenir més dades per tal de poder evidenciar aquesta no linealitat de forma generalitzada. El model lineal fou comparat amb altres models estadístics com l'ajust amb funcions lineals definides per trams; la regressió logística (logit); la regressió quadràtica; i els models additius generalitzats (GAM).
Altres autors han analitzat els models multinivell com a alternativa al model de regressió lineal per a estimar l'accés a fonts d'aigua i a instal·lacions de sanejament millorat (Wolf, Bonjour, & Prüss-Ustün, 2013). Així mateix, la natura composicional de les dades amb les quals treballem fa necessari incloure en aquesta reflexió l'anàlisi composicional com una de les metodologies alternatives a la del JMP.

A continuació s'enumeren de forma somera aquests mètodes alternatius, incloent l'anàlisi composicional, que es desenvoluparà en profunditat en apartats posteriors (apartats 7.2 i 8).

5.1 Funcions lineals definides per trams

Una funció definida per trams és una funció la definició de la qual canvia segons el valor de la variable independent. Al tractar-se d'una funció lineal, cadascun dels trams serà lineal.

Formalment, una funció real *f* (definida a trams) d'una variable real *x* és la relació la definició de la qual està donada per diversos conjunts disjunts del seu domini (subdominis).

Aquest ajust suposa dividir el període d'anàlisi de dades en subperíodes i realitzar diferents ajustos lineals per a cadascun d'ells. La funció ha de ser, evidentment, contínua. Si al punt d'intersecció (comú) de les rectes l'anomenem node, la clau radica en determinar el nombre de nodes necessaris (nombre de trams que realitzem) i la seua posició.

A la figura 4 podem observar un exemple comparatiu de l'ajust amb funcions lineals definides per trams i el mètode del JMP pel que fa a la cobertura d'instal·lacions millorades d'aigua en el cas de zones urbanes en Paraguai i instal·lacions canalitzades en les llars a les zones Rurals en El Salvador.

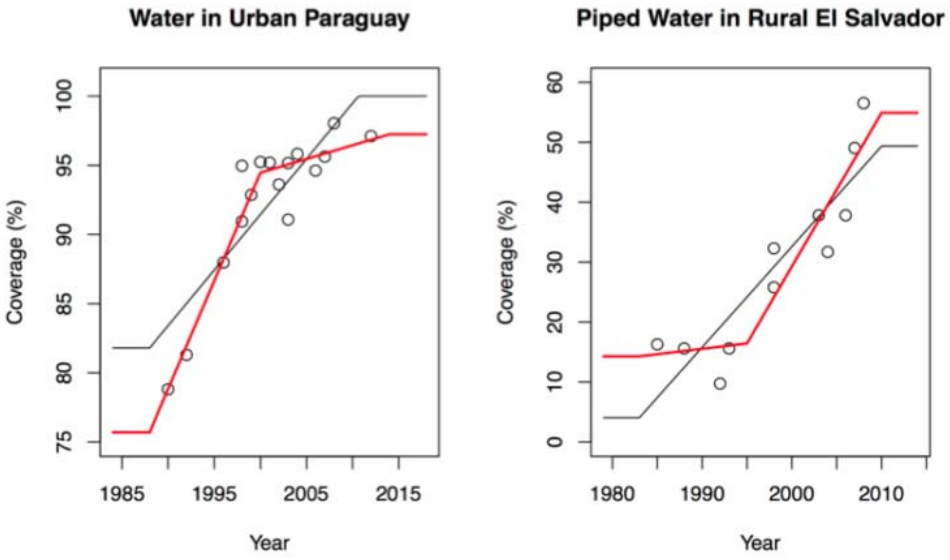


Figura 4. Exemple comparatiu model d'ajust mitjançant regressió lineal definida a trams vs RLS (JMP). Font: (WHO & UNICEF, 2014).

Com podem veure, el mètode alternatiu proposat contempla una funció continua formada per dues regressions lineals front a una única regressió al cas del JMP. En ambdós mètodes, es força una pendent nul.la per a les estimacions immediatament anteriors a la primera observació i immediatament posteriors a l'última, com hem vist a l'apartat 4.3 i a la figura 3.

5.2 Regressió logística (logit)

La regressió logística, que consisteix en la transformació de la variable dependent utilitzant una funció logit, dóna lloc a una corba en forma de "S". Alguns autors suggereixen que aquesta podria correspondre a la tendència general per a aigua i sanejament però tanmateix, no pot afirmar-se açò per al conjunt de sèries de dades del JMP. El model funciona bé als casos de saturació (alentiment en les tendències quan els percentatges de cobertura s'acosten al 100%) però si es produeix una acceleració per nivells de cobertura intermedis, la tendència és lineal (WHO & UNICEF, 2014).

Un exemple comparatiu d'una regressió logística amb el model de RLS del JMP és el de la figura 5, on s'aprecien els ajustos per a les dades d'accés a una font millorada en entorn Urbà en Paraguai i a instal·lacions de sanejament millorades també en entorn urbà, ara a Tanzània.

Com podem apreciar, la regressió logística té forma corba. L'extrapolació d'estimacions a l'inici i al final de la sèrie (immediatament abans del primer valor observat i després de l'últim) es força manualment, imposant pendent nul.la de la mateixa manera que per al mètode del JMP.

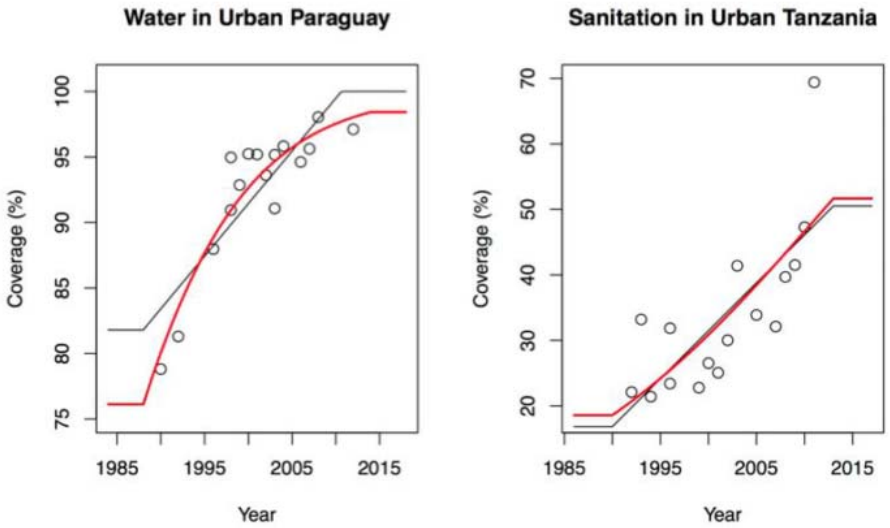


Figura 5. Exemple comparatiu model d'ajust mitjançant regressió logística vs RLS (JMP).
Font: (WHO & UNICEF, 2014).

5.3 Regressió quadràtica

La regressió quadràtica consisteix en ajustar al conjunt de valors una funció polinòmica de segon grau.

Al gràfic següent (figura 6) podem observar un exemple comparatiu de l'ajust quadràtic front la RLS utilitzada pel JMP. Les dades de partida són les mateixes que per al cas de la figura 5, és a dir, accés a font millorada d'aigua en entorn urbà en Paraguai i a una instal·lació millorada de sanejament en entorn urbà en Tanzània.

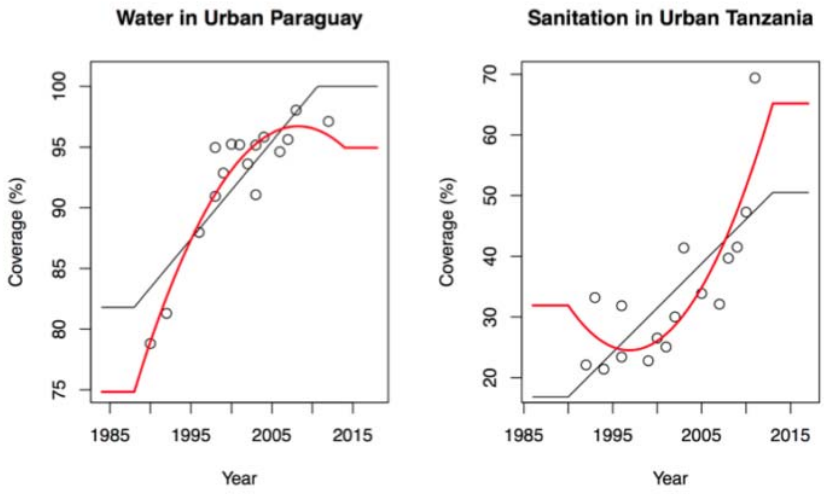


Figura 6. Exemple comparatiu model d'ajust quadràtic vs RLS (JMP).
Font: (WHO & UNICEF, 2014).

L'anàlisi de les regressions quadràtiques realitzat per Fuller et al. citat en WHO & UNICEF, 2014 assenyala que aquestes funcions forcen l'existència d'un punt d'inflexió, el que no es correspon als patrons observats a les dades del JMP. Els ajustos quadràtics poden conduir a resultats extrems per a sèries amb pocs punts (n<5) i, no es considera com una alternativa viable al mètode del JMP.

5.4 Models additius generalitzats (GAM)

Els models additius generalitzats són una aproximació que pot utilitzar-se per representar una corba amb canvis de curvatura suaus. Aquests models no tenen una funció de forma específica sinó que més bé suposen penalitzar corbes amb molta curvatura a partir de la sèrie de dades disponible. Quan existeixen pocs punts, els algoritmes acaben utilitzant un model simple que tendeix al model lineal a mesura que el nombre de punts disponible es fa més petit (WHO & UNICEF, 2014).

A la figura 7 es mostra la comparació entre l'ajust mitjançant un model GAM i el mètode del JMP per a les dades de cobertura en l'accés a una font d'aigua millorada en entorn urbà a Paraguai i d'accés a una instal.lació de sanejament millorada en entorn urbà a Tanzània. Com pot apreciar-se, la predicció als extrems es torna a forçar manualment per imposar una pendent nul.la, de la mateixa manera que ho fa el mètode del JMP.

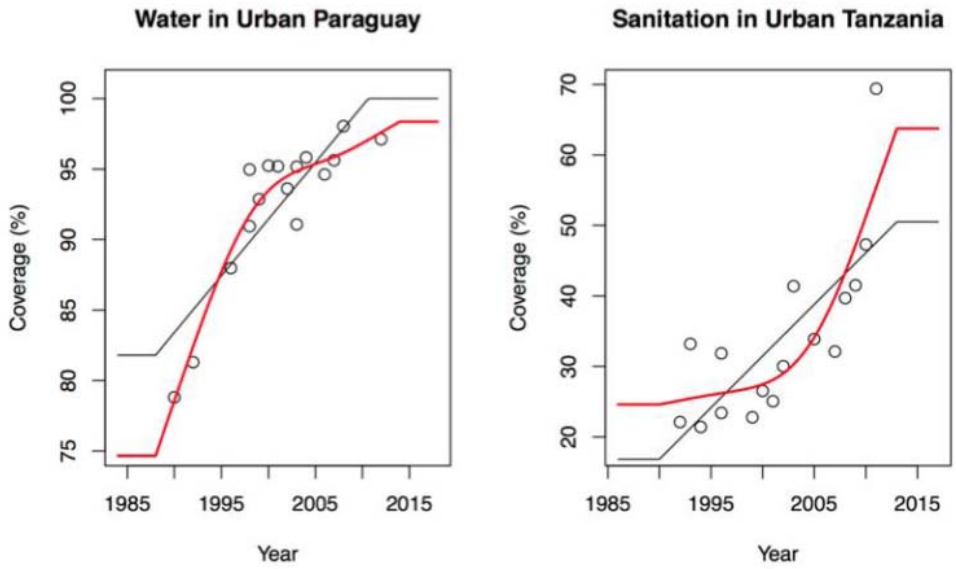


Figura 7. Exemple comparatiu model d'ajust quadràtic vs RLS (JMP).
Font: (WHO & UNICEF, 2014).

5.5 Models Multinivell (MLM)

Els models multinivell permeten corregir els paràmetres que defineixen l'ajust lineal per a cada país, segons si les informacions provinents de la sèrie de dades del propi país es consideren fiables o no. En la pràctica, el que fan és assumir que l'evolució de les variables als països es produeix d'acord amb la tendència mitjana regional al cas en que la informació sobre aquesta tendència a partir de les dades del país siga escassa o nul.la (Wolf et al., 2013).

Així, al cas que siguen fiables (és a dir, la sèrie de punts disponible és gran i la variància entre ells és petita) s'aplica el model lineal simple, és a dir, la recta de regressió per a l'ajust es determina únicament a partir de la sèrie de dades disponibles del país.

En el cas contrari (pocs punts i/o alta variabilitat entre ells), tot i que núvol de punts és utilitzat per determinar l'ordenada en l'origen de la recta de regressió, la pendent (que representa la tendència) s'assimila a la pendent mitjana regional (Wolf et al., 2013).

Wolf et al. (2013) han comparat la utilització d'aquests models amb la sèrie de dades del JMP. Específicament citen com a un dels seus avantatges front la RLS, la disposició de sèries de dades d'estimacions contínues per a tots els països durant tot el període dels ODM, inclús al cas on les dades disponibles són escasses.

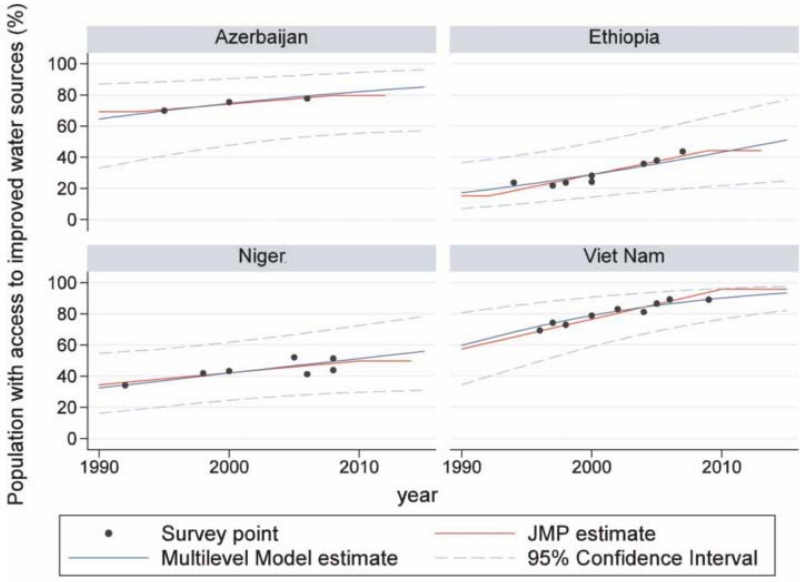


Figura 8. Exemple comparatiu ajust mitjançant MLM vs RLS (JMP).
Font: (Wolf et al., 2013).

5.6 Anàlisi composicional

Per anàlisi composicional ens referim a l'anàlisi estadística de dades de tipus composicional. Les dades composicionals són dades que descriuen quantitativament les distintes parts d'un tot i ens donen únicament informació relativa entre les seues components (Egozcue & Pawlowsky-Glahn, 2011a) (Pawlowsky-Glahn et al., 2015).

Habitualment s'expressen en una forma tancada (o clausurada), el que significa que la seva suma és constant. Aquest és el cas de dades com les del JMP, que representen percentatges (o proporcions respecte a la unitat) de població utilitzant un tipus determinat de font d'aigua o d'instal·lació de sanejament.

Les dades composicionals tenen particularitats específiques i propietats numèriques que condicionen qualsevol anàlisi estadística que es realitze sobre elles (Pawlowsky-Glahn & Egozcue, 2006). La conseqüència essencial d'aquestes propietats és que les tècniques habituals d'anàlisi estadística, ideades per a variables aleatòries sense restriccions, no poden ser utilitzades per analitzar dades composicionals directament o en "cru", és a dir, sense un tractament matemàtic previ (Pawlowsky-Glahn & Egozcue, 2006) (John Aitchison & Egozcue, 2005).

El fet de no tenir en compte el caràcter composicional de les dades pot conduir a valoracions errònies, invalidant la interpretació dels resultats de l'anàlisi estadística clàssica sobre aquestes. A pesar que des de fa molt de temps s'està advertint d'aquest fet, continua utilitzant-se l'anàlisi estadística "clàssica" per aquest tipus de dades. (Egozcue & Pawlowsky-Glahn, 2011a) (Pawlowsky-Glahn & Egozcue, 2006).

Així doncs, donat que les dades del JMP d'accés a aigua i sanejament són, com ja s'ha dit, dades composicionals, l'anàlisi del JMP que suposa l'aplicació dels mètodes estadístics clàssics (regressió lineal simple) no resulta adequat i les interpretacions que se'n deriven d'aquesta anàlisi podrien ser errònies. Mètodes alternatius com el de les funcions lineals definides per trossos presentarien el mateix problema donat que es segueix aplicant l'estadística clàssica (tot i que per trams distints) a una sèrie de dades que no admet aquest tipus de tractament.

Com indiquen alguns autors, el problema clau d'aplicar l'estadística clàssica sobre dades composicionals és la incertesa que acompanya els resultats: no és possible distingir entre els efectes anomenats espuris[1] causats per la restricció de que la suma siga constant i els efectes atribuïbles a processos naturals (Pawlowsky-Glahn & Egozcue, 2006). Per aquesta raó resulta fonamental aplicar la metodologia adequada per a aquest tipus de dades, i comparar els resultats obtinguts amb els derivats de l'anàlisi amb l'estadística clàssica (mètode del JMP) per intentar determinar en quin grau els efectes espuris dominen el procés i impedeixen interpretar correctament aquests resultats.

Aquest és, precisament, l'objecte d'aquesta tesina: l'anàlisi i comparació dels resultats obtinguts amb el tractament estadístic clàssic (erroni, com hem dit) i amb les tècniques composicionals.

A més, el tractament d'aquestes dades com composicionals permet garantir una sistematització del procés d'anàlisi, sense manipular les dades extrapolades amb el model utilitzat per forçar el seu valor, com fa el JMP –i la resta de mètodes presentats- al cas de les projeccions abans i després del primer i últim cens o mostreig del qual es disposa.

A l'apartat 7, de Metodologia, s'aprofundirà en la metodologia de l'anàlisi composicional. Tanmateix, de manera resumida i a mode d'introducció, podem dir que l'anàlisi de dades composicionals pot reduir-se als tres passos següents (Egozcue & Pawlowsky-Glahn, 2011a):

1. La transformació de les dades a unes coordenades distintes (coordenades tipus log-quocient);
2. L'anàlisi estadística (clàssica) de les dades en coordenades "transformades";
3. La transformació inversa o recuperació de les coordenades inicials (expressió de les dades en forma de composicions) i la interpretació dels resultats obtinguts.

[1] La relació espúria és una relació matemàtica en la qual dos successos no tenen connexió lògica, tot i que existeix una correlació estadística, degut a l'existència d'un tercer factor no considerat. Aquesta relació dóna la impressió de la existència d'un vincle apreciable entre els dos grups però què és invàlid quan s'examina objectivament

6 DADES DE BASE

6.1 Fitxers de dades originals

La base de dades de partida ha estat proporcionada directament per a la seua anàlisi pel JMP, durant el primer semestre de l'any 2015. Es tracta del conjunt de dades que ha servit de base al JMP per elaborar l'Informe sobre el progrés en matèria d'Aigua i Sanejament en relació als ODM (actualització del 2015) (WHO/UNICEF, 2015).

Més concretament, la informació es troba enregistrada en dos fitxers, amb format .csv, un per a les dades d'aigua (Water ladder 2015.csv) i altre per a les dades de sanejament (Sanitation ladder 2015.csv).

El fitxer d'aigua és una taula que s'estructura amb les següents sis columnes:

Country	Survey_code	Year	Rural	Urban	Type

On:

- "Country": és el nom del país al qual corresponen les dades de la fila
- "Survey_code": és el codi que identifica el sondeig o el cens del qual prové la dada
- "Year": és l'any al qual correspon el sondeig/cens realitzat
- "Rural": és el percentatge de població rural que s'abasteix d'una font d'aigua del tipus "type"
- "Urban": és el percentatge de població urbana que s'abasteix d'una font d'aigua del tipus "type"
- "Type": correspon al tipus de font, d'acord amb l'escala d'aigua i la nomenclatura del JMP. Pot prendre els següents valors:
  - "S", per "Surface". Correspon a l'aigua d'origen superficial
  - "p", per "piped". Es refereix a "canalitzada", entesa com canalitzada dins de la propietat de la llar. És un tipus particular de font millorada.
  - "I", per "improved". Es refereix a font millorada d'aigua, comprenent les canalitzades també.

El fitxer te un total de 4280 files. Cadascuna d'aquestes files correspon a les dades de cobertura per al tipus de font del que es tracte (a la columna "type"), per al sondeig realitzat en un país i un any determinat.

El fitxer de Sanejament és una taula estructurada de la mateixa manera que l'anterior. Així, es composa de les 6 columnes següents:

Country	Survey_code	Year	Rural	Urban	Type

I respecte a l'anterior, el que canvia és la variable "type". En aquest cas pot ser:

- "Sr", de Shared, que correspon als valors d'instal·lació millorada compartida.
- "I", per "Improved", o instal·lació de sanejament millorada (compartida o no compartida).
- "Od", per "Open Defecation" o defecació a l'aire lliure.

Aquest fitxer te un total de 3110 files que corresponen als valors observats per als distints "type" en la totalitat de mostrejos realitzats a nivell mundial.

6.2 Reorganització de dades i tractament

Per poder realitzar l'anàlisi composicional de les dades, necessitem que aquestes dades siguen efectivament composicionals. D'acord amb la definició de dades composicionals cal ser capaços d'identificar el "tot" i les distintes parts d'aquest.

Al nostre cas, de la mateixa forma que ho fa el JMP, distingirem entre l'accés a l'aigua i l'accés al sanejament i separarem el conjunt en dues subpoblacions: corresponents a l'entorn rural i a l'entorn Urbà.

Així, del conjunt de dades de cada fitxer original (aigua i sanejament) s'han separat les dades corresponents a entorn urbà de les de l'entorn rural, per analitzar-les separadament. Totes les files de cada sèrie de dades (aigua rural, aigua urbana, sanejament rural i sanejament urbà) amb algun valor "NA" han estat eliminades per evitar problemes en l'anàlisi.

Les dades corresponents a accés a l'aigua (en medi rural per un costat i urbà per altre) i accés a instal·lacions de sanejament (rural/urbà) són dades, com hem dit, de tipus composicional. A cadascuna d'aquestes (rural/urbà), el "tot" és la proporció completa (100%) de la població.

Al cas de l'accés a l'aigua, la composició pot descompondre's en dos parts (proporció de gent respecte del total amb accés a una font "millorada" – *Improved*- d'aigua i proporció de gent respecte del total amb accés a una font "no millorada" –*Unimproved*- d'aigua).

Aquestes dues parts, alhora, poden descompondre's en dos més cadascuna: canalitzat en la parcel.la –*Piped on premises*- i altres tipus –*Other Improved*- , en el cas dels "millorats" i font d'aigua superficial – *Surface*- i altres tipus –*Other Unimproved*-, al cas dels "no millorats".

L'esquema següent (figura 9) il·lustra les distintes components de l'anàlisi. El primer nivell (millorada/no millorada) representa una composició amb 2 parts i el segon nivell una composició amb 4 parts.

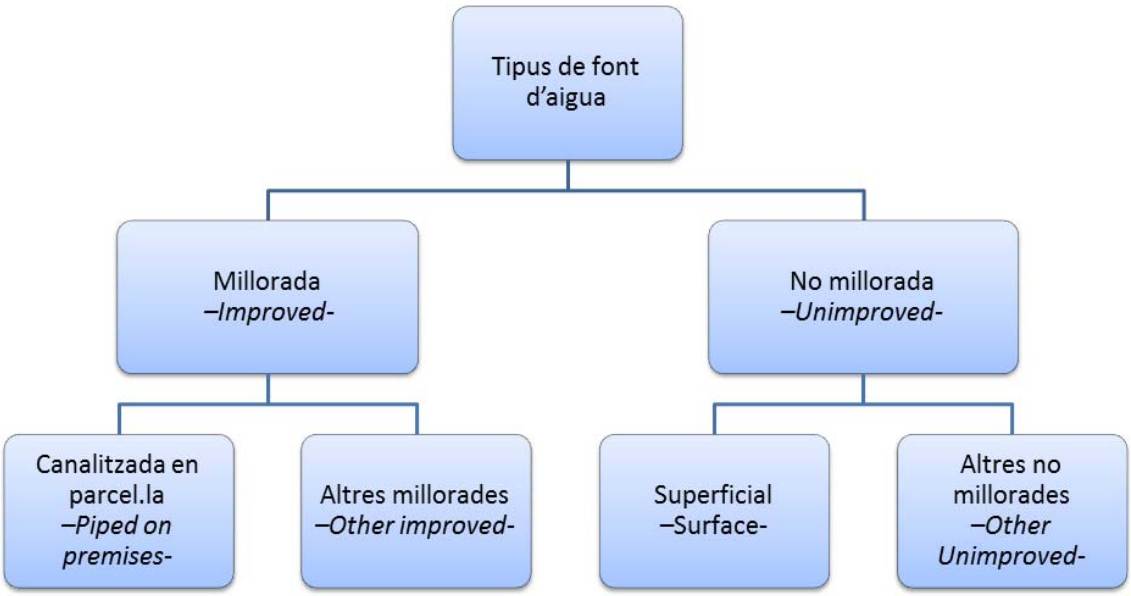


Figura 9. Components composicionals de les dades d'aigua. Font: elaboració pròpia

Donat que les dades recollides a la sèrie de dades del JMP no corresponen directament a les distintes components de l'anàlisi composicional que acabem de veure, és necessari reorganitzar les dades originals prèviament al seu tractament.

Així, a partir de les sèries de dades originals (descrites a l'apartat 6.1), les components composicionals pel que fa a l'accés a una font d'aigua potable s'han obtingut de la següent forma:

- Millorada (*Improved*), que l'anomenarem "I": és una observació recollida directament a la sèrie original.
- No millorada (*Unimproved*, "U"): l'obtenim a partir de l'anterior, com:
  Unimproved (U) =1 - Improved (I)
- El total dels "*Improved*" poden descompondre's en:
  - P (*piped on premises*): és una observació recollida a les dades del JMP.
  - OI (*Other Improved*) = I (Improved)-p
- Dels "*Unimproved*",
  - S (*surface*): és una dada que prové directament de la sèrie original.
  - OU (*Other Unimproved*) = U (*unimproved*) – S (*Surface*)

Per al cas del sanejament, les dades originals s'han separat en dos grans grups, en funció de l'entorn on s'han recollit (rural i urbà) per analitzar-les independentment. A dintre de cadascun d'aquests dos grups, al conjunt de dades corresponent al sanejament podem identificar dos composicions de la mateixa manera que hem fet per a les dades d'aigua.

Així, podem identificar una primera composició de dues parts (instal·lacions millorades *–Improved-* i no millorades *–Unimproved-*) i una segona composició en quatre parts, resultat de descompondre cadascuna de les dos parts anteriors en altres dos: Millorades compartides (*Improved Shared –IS-*) i

no compartides (*Improved not Shared –InS-*) i defecació a l'aire lliure (*Open Defecation – OD-*) i altre tipus de no millorades (*Other Unimproved – OU-*).

La figura 10 mostra les distintes parts de l'anàlisi:

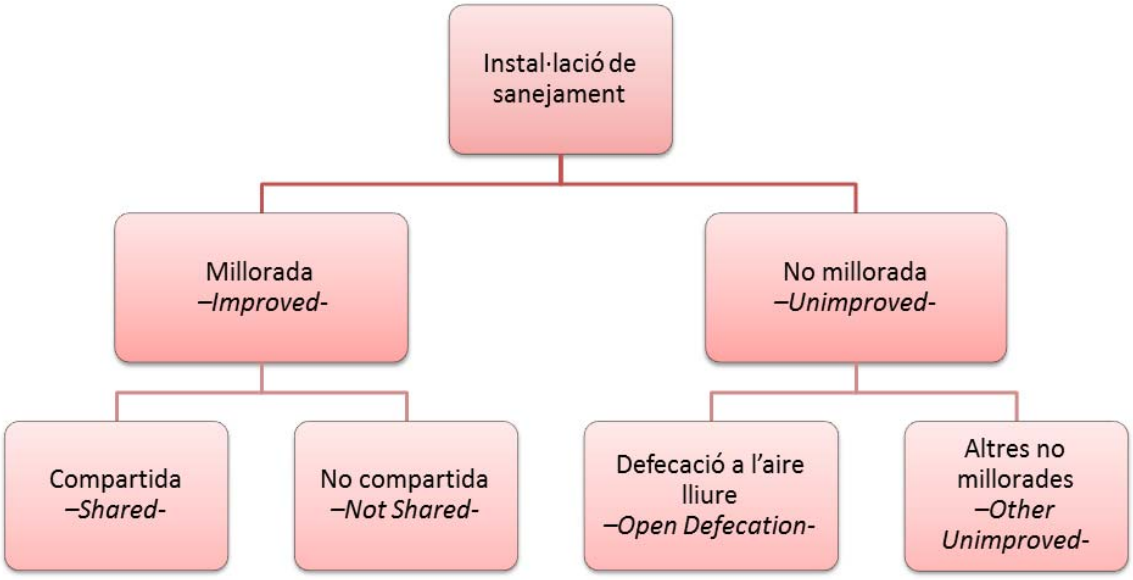


Figura 10. Components composicionals de les dades de sanejament. Font: elaboració pròpia

L'obtenció de les distintes components s'ha realitzat de la següent forma:

- I (*Improved*): aquesta dada prové directament de la sèrie de dades original del JMP
- U (*Unimproved*): l'obtenim com U=1-I (Improved)
- OD (*Open Defecation*): dada recollida a la sèrie original del JMP
- OU (*Other Unimproved*) = U – OD
- IS (*Improved Shared*): les dades de les quals es disposa són de proporcions (en tant per 1) d'instal.lacions millorades compartides respecte del total de millorades (*sr*). D'aquesta manera, el valor de IS pot obtenir-se com:

  $$IS = sr * I$$

  El problema en aquest cas rau en que la sèrie de dades no és completa, és a dir, el valor de *sr* no ha estat observat en tots els sondejos realitzats. El valor finalment utilitzat de *sr* pera cada país és el valor mitjà dels valors disponibles per aquest. A la sèrie de dades del JMP, com ja s'ha assenyalat a l'apartat 4.3, existeixen països on malauradament no hi ha cap valor observat de *sr* en la sèrie de dades. Per tal de poder sistematitzar l'anàlisi, s'ha prescindit de les dades corresponents als països sense cap informació en relació al percentatge d'instal.lacions millorades de sanejament d'ús compartit.
- InS (*Improved not Shared*) = I – IS

6.3 Sèrie de dades de treball

La metodologia seguida per obtenir els fitxers amb les sèries de dades de partida per al seu anàlisi ha estat descrita parcialment a l'apartat anterior. Aquesta, de manera completa, es descriu a continuació:

1. Dels fitxers originals de dades del JMP (dos fitxers de tipus .csv, un amb les dades d'aigua i altre amb les de sanejament) hem generat quatre taules (fitxers): dos per a aigua i dos per a sanejament. Cadascun dels dos d'un tipus determinat correspon a les dades referents a les subpoblacions rurals i urbanes.
2. De les quatre sèries de dades anteriors, aquelles files amb algun valor inexistent ("NA") són eliminades.
3. Es reorganitzen les dades a cadascuna de les quatre sèries de forma que cada fila correspon a una observació (corresponent a un cens o mostreig realitzat) i organitzant en columnes les observacions corresponents al distints tipus d'instal·lació (d'aigua o sanejament segons siga el cas).
4. Es realitzen les operacions bàsiques amb les variables originals per obtenir els distints factors per a l'anàlisi composicional.
   Els fitxers resultants seran sèrie de dades en format .csv amb la següent estructura:
   a. Aigua (una sèrie per a entorn URBÀ i altre per entorn RURAL)

Country	*Survey_code*	*Year*	*I*	*U*	*P*	*OI*	*S*	*OU*
			I_j	$U_j=1- i_j$	p_j	$OI_j=I_j-p_j$	s_j	$OU_j=U_j-s_j$

   b. Sanejament

Country	*Survey_code*	*Year*	*I*	*U*	*IS*	*InS*	*OD*	*OU*
			I_j	$U_j=1- I_j$	$IS_j=sr_k^2*I_j$	$InS_j=I_j-IS_j$	OD_j	$OU_j=U_j-OD_j$

L'esquema de la figura 11 il·lustra la metodologia seguida fins aconseguir les sèries de dades de partida de l'anàlisi composicional.

[2] El valor de sr_k és el valor mitjà del percentatge d'infraestructura de sanejament millorada compartida del país "k", al qual correspon el cens realitzat (cens "j")

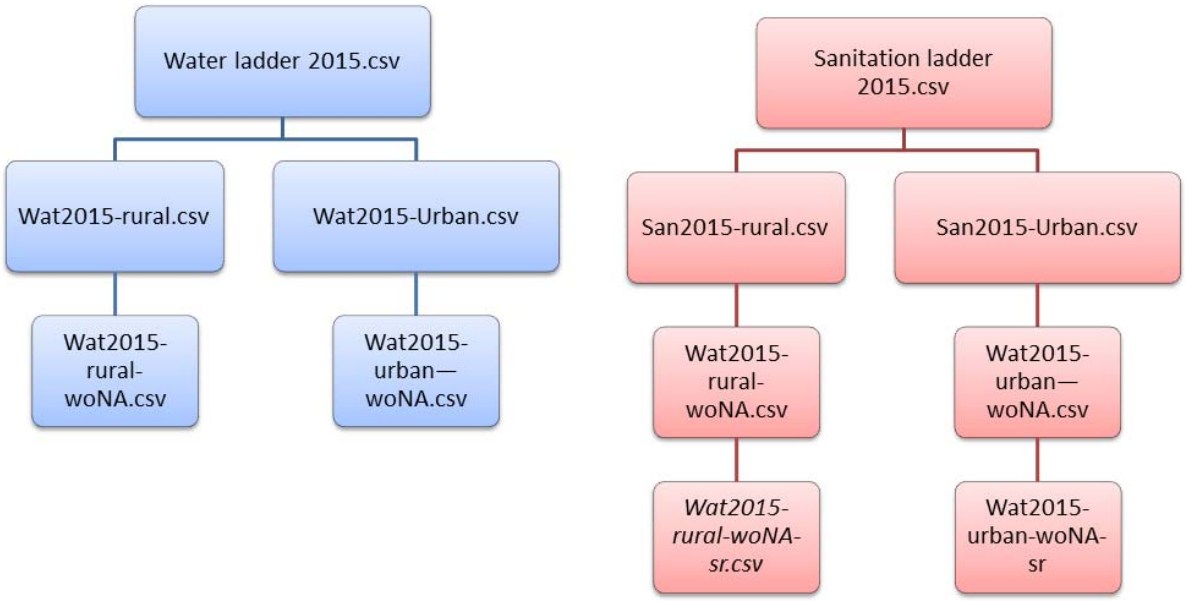


Figura 11. Procés de construcció de les sèries de dades de treball. Esquema

Del conjunt de països del llistat complet del JMP nosaltres ens anem a centrar en la regió de l'Àfrica Subsahariana (SSA per les seues sigles en anglés). Aquesta és una de les 10 regions identificades per Nacions Unides per tal de monitorar el compliment dels Objectius del Mil.leni[3], i que el JMP pren com a referència per als seus estudis.

Precisament hem agafat aquesta regió com a referència ja que la immensa majoria del conjunt de països classificats per Nacions Unides com els menys desenvolupats ("*Least Developed Countries*", o *LDC*, segons la terminologia anglesa) es troben en ella (figura 12).

És en aquests tipus de països on la situació de partida respecte a l'accés a aigua i a instal·lacions de sanejament era més precària (taxes de cobertura pel que fa a l'ús de fonts o instal·lacions millorades més baixes) al 1990 (any de referència per a la meta 7c) i on intuïtivament podem suposar que s'han realitzat progressos més significatius.

[3] Segons aquesta classificació, el món es divideix en les 10 regions següents: Àfrica Subsahariana (1), Nord d'Àfrica (2), Àfrica de l'Est (3), Sud d'Àsia (4), Sud Est Asiàtic (5), Àsia de l'Oest (6), Oceania (7), Amèrica Llatina i el Carib (8), Caucas i Àsia Central (9) i una altra regió (10), que no respon a criteris geogràfics, sinó que està constituïda per països dins de regions (en sentit geogràfic) que es consideren desenvolupades (d'acord amb l'IDH del país).

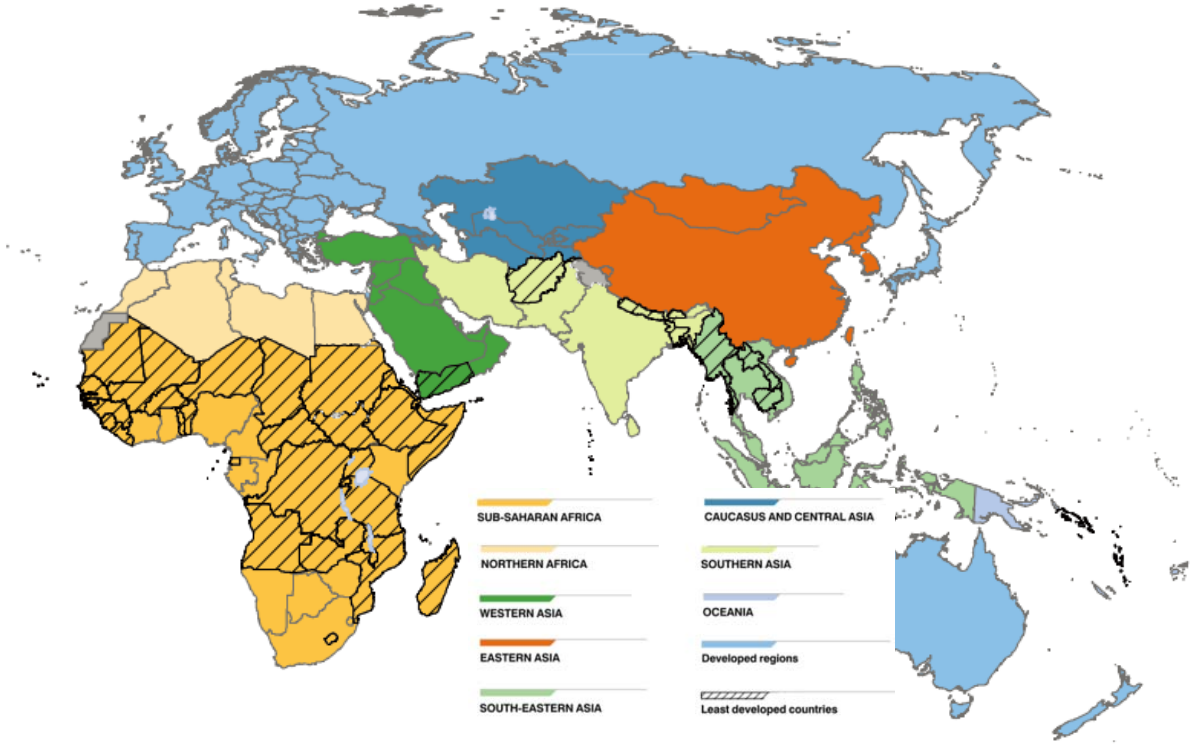
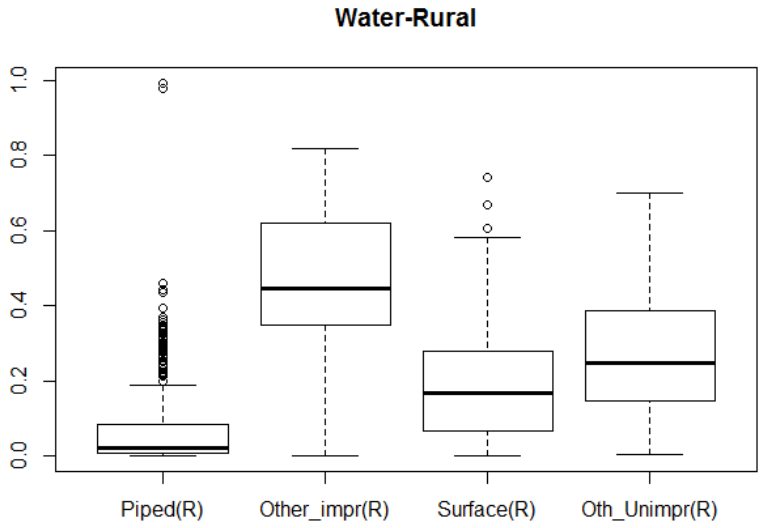


Figura 12. Mapa de divisió regional utilitzat pel JMP (extracte). Font: (WHO/UNICEF, 2015)

La representació dels diagrames de caixes (boxplots) del conjunt d'observacions disponibles per a tots els països de la regió SSA per a cadascuna de les quatre sèrie de dades (WR, WU, SR i SU), permet fer-nos una idea inicial sobre el tipus i la distribució de dades que tenim.

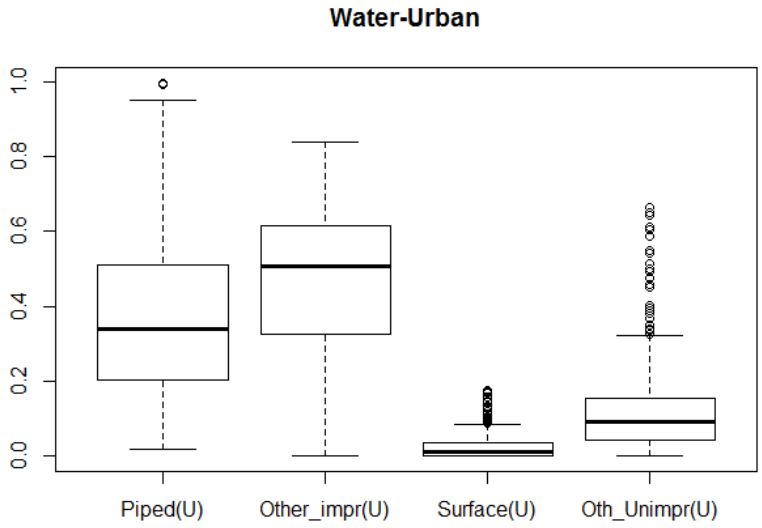
Els diagrames de caixes permeten veure clarament el valor mitjà del conjunt per a cada part (línia horitzontal més gruixuda a l'interior de la caixa), la distribució de valors respecte a aquest i detectar valors anòmals (aquells que es troben fora de les línies horitzontals extremes).

En el cas d'Aigua en entorn Rural, el gràfic 2 ens permet observar que en entorn Rural el menys habitual és tenir una instal·lació canalitzada en el domicili (la majoria de valors es troben per baix del 20% tot i que tenim bastants valors atípics) i que predominen els altres tipus de fonts millorades front les altres.



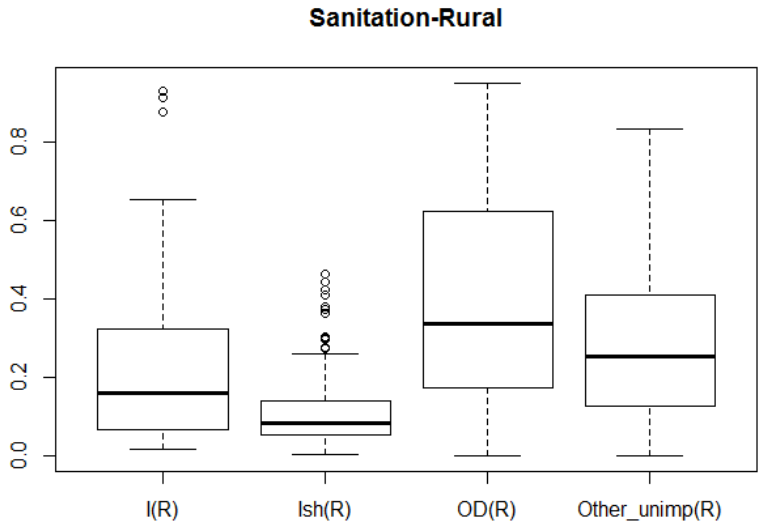
Gràfic 2. Representació en diagrama de caixes (box-plot) de la sèrie de dades composicionals d'accés a una font d'aigua en entorn Rural. Font: Elaboració pròpia

Al cas d'accés a font d'aigua en entorn urbà (gràfic 3), com era d'esperar, tenim que el menys comú és abastir-se d'una font superficial (pou a cel obert, curs d'aigua, etc.) i que predominen les fonts millorades front les no millorades.

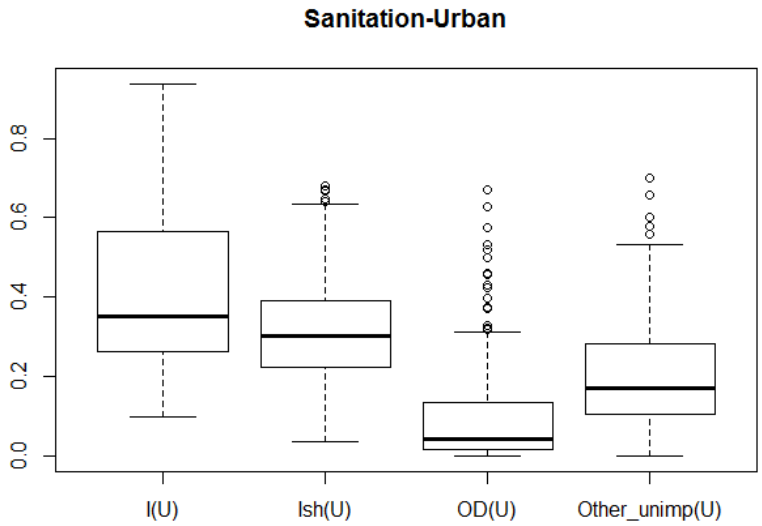


Gràfic 3. Representació en diagrama de caixes (box-plot) de la sèrie de dades composicionals d'accés a una font d'aigua en entorn Urbà. Font : elaboració pròpia

Per a les dades de sanejament, l'estructura de les dades és tal i com podia preveure's: al cas d'entorn Rural, les instal·lacions no millorades són més comunes que les millorades, amb una importància notable de la defecació a l'aire lliure (gràfic 4); i al cas de l'entorn Urbà, on les instal·lacions millorades són més comunes tot i que en gran proporció d'ús compartit i la pràctica de la defecació a l'aire lliure és poc habitual (gràfic 5).



Gràfic 4. Representació en diagrama de caixes (box-plot) de la sèrie de dades composicionals d'accés a una instal·lació de sanejament en entorn Rural. Font : elaboració pròpia



Gràfic 5. Representació en diagrama de caixes (box-plot) de la sèrie de dades composicionals d'accés a una instal·lació de sanejament en entorn Urbà. Font : elaboració pròpia

7 METODOLOGIA

A aquesta secció s'exposa de manera detallada la metodologia general que permet ajustar un model de regressió lineal simple al conjunt de dades disponibles (de manera directa), i la metodologia pròpia de l'anàlisi composicional. A l'apartat 8 es particularitzarà la metodologia exposada per al cas de les dades del JMP per un país i un entorn (rural o urbà) concret.

Cal assenyalar que estem assumint que la metodologia utilitzada pel JMP correspon al model de regressió lineal simple, sense tenir en compte, a nivell metodològic, els ajustos realitzats per evitar valors fora de rang (veure 4.3). A mesura que es presenten els resultats (apartat 8) es discutirà sobre la qüestió, fent una menció específica en l'apartat 8.1.1.5.

7.1 Model de regressió lineal simple (RLS)

El model de regressió lineal simple expressa la relació entre una variable resposta "Y" i una variable predictora "X" de la següent forma:

$$Y = \beta_0 + \beta_1 \cdot X + \varepsilon \qquad \text{Equació (1)}$$

On:

- β_0 i β_1 són dues constants que representen els paràmetres o els coeficients de regressió del model
- ε és l'error o soroll aleatori.

D'aquesta manera estem acceptant que per a les dades d'estudi, l'equació lineal anterior ens dóna una aproximació acceptable de la relació real entre X i Y. En altres paraules, estem assumint que Y és aproximadament una funció lineal de X, i ε mesura les desviacions o discrepàncies en aquesta aproximació (Montgomery, Peck, & Vining, 2015).

El coeficient β_1 ("pendent"), pot interpretar-se com el canvi en Y per unitat de canvi en X. El coeficient β_0 (anomenat "ordenada a l'origen"), és el valor predit d'Y quan X=0.

Cada observació (y_i) pot expressar-se com:

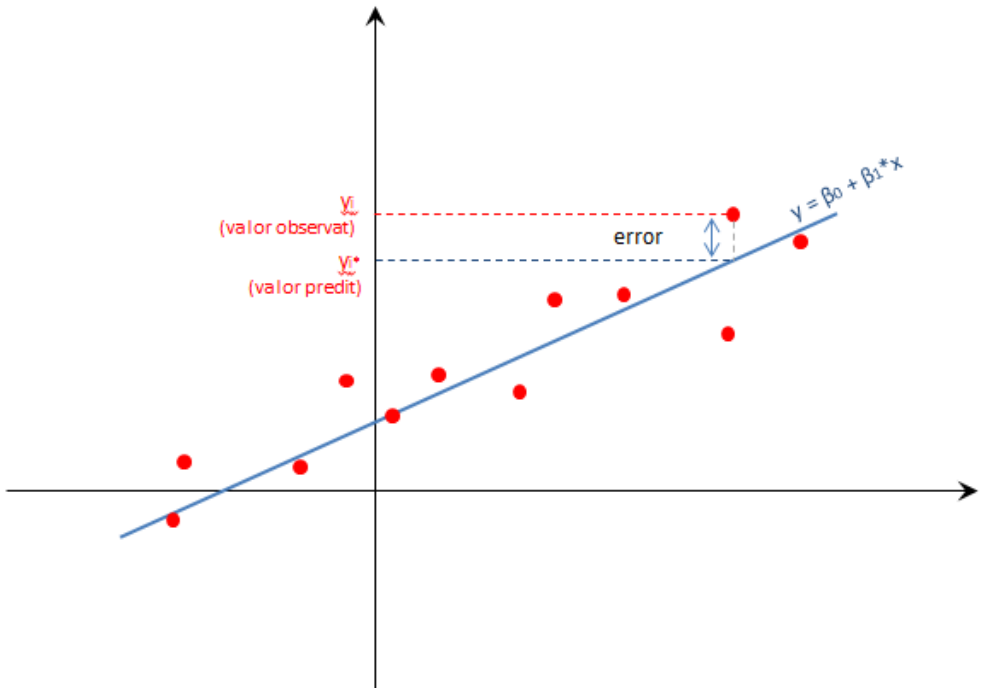
$$y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \qquad \text{Equació (2)}$$

L'objectiu per tal de definir correctament el model és determinar els valors dels paràmetres d'aquest (β_0 i β_1). El mètode normalment utilitzat per fer-ho és el mètode dels mínims quadrats (Zou, Tuncali, & Silverman, 2003). Els paràmetres obtinguts a partir d'aquest mètode minimitzen la suma del quadrat de les distàncies verticals entre el valor observat i el valor estimat, és a dir, la suma del quadrat dels errors ε_i.

Si y_i^* és el valor d'y_i estimat pel model de regressió lineal, l'error o distància vertical entre aquest i el valor realment observat (gràfic 6), ve donat per:

$$e_i = y_i - y_i^* \qquad \text{Equació (3)}$$
; amb i=1, 2, ..., n

Aquestes distàncies verticals s'anomenen els residus ordinaris de mínims quadrats. Una de les propietats d'aquests residus és que la seua suma és zero.



Gràfic 6. Regressió lineal simple. Ajust per mínims quadrats. Font: elaboració pròpia

Una vegada realitzat l'ajust no sols ens interessa conèixer l'existència d'una relació lineal entre les dues variables (depenent i independent) sinó també poder quantificar la qualitat de l'ajust del model a les dades originals (Montgomery et al., 2015).

Aquesta quantificació ens la dóna el coeficient de determinació (R^2), que per al cas de la regressió lineal simple és igual al quadrat del coeficient de correlació o coeficient de Pearson, què és el quocient entre la covariància de (X,Y) i el producte de les desviacions típiques de x i de y.

$$r = \frac{\sigma_{XY}}{\sigma_X \cdot \sigma_Y} \qquad \text{Equació (4)}$$

Així, R^2 és un indicador de la bondat de l'ajust, que pren valors entre 0 i 1. Els valors pròxims a 1 ens indiquen una millor qualitat de l'ajust.

Tanmateix, per jutjar la idoneïtat del model per a les dades de les quals es disposen (en definitiva, per jutjar la bondat de l'ajust) cal comprovar també que els supòsits sobre els quals es basa el model són certs. En cas que no siga així, caldria plantejar-nos en quina mesura el model lineal és adequat per a reproduir el comportament de les dades disponibles.

Aquests supòsits (o hipòtesis) del model lineal fan referència generalment als residus, és a dir, a la diferència entre els valors observats realment i aquells predits pel model.

Es tracta dels següents supòsits:

1. Linealitat: La relació entre la variable depenent i independent ha de ser lineal.
   Per tal de jutjar si aquest supòsit es compleix,:
   - Podem observar visualment si la distribució de valors observats front la variable depenent és "aproximadament" lineal.
   - Analitzar la representació gràfica dels Residus front als Valors Predits o estimats pel model lineal. Aquest gràfic permet observar si la distribució dels residus respecte als valors predits respon a patrons no lineals.
     Podria donar-se el cas on existira una relació no lineal entre la variable predictora i una variable externa, que quedaria recollida a aquest gràfic i que el model (lineal) no estaria tenint en compte.
     Si els errors es distribueixen aleatòriament al voltant del zero, sense reproduir cap patró, tenim un indicador de la inexistència de relacions no lineals.

2. Independència: els errors (residus) han de ser independents tant dels valors predits com entre ells.

3. Homoscedasticitat. Aquest supòsit exigeix que per a tot el recorregut de la variable independent, la variància de l'error siga constant. Per comprovar la homoscedasticitat representarem el gràfic de la variància dels residus front Valors Predits (gràfic comunament anomenat "Scale-Location").

4. Normalitat dels residus, és a dir, els errors segueixen una distribució Normal. El diagnòstic d'aquest supòsit es farà a partir del diagrama Quantil-Quantil (gràfic "Normal Q-Q"). La hipòtesi de normalitat no es pot rebutjar si el núvol de punts es distribueix linealment a aquest gràfic.

Altre gràfic interessant és la representació dels residus front l'apalancament, anomenat "Residuals vs Leverage", que representa la distància de Cook de cada observació. Aquesta és una mesura de la influència d'aquestes observacions sobre el model. Així, valors superiors a la unitat representen observacions clau o molt influents, que són capaces de condicionar el model, forçant-lo de tal forma que la simple eliminació d'aquesta observació faria que els resultats foren completament distints.

7.2 Anàlisi composicional

Com hem assenyalat a l'apartat 5.6, definim les dades composicionals com dades que descriuen quantitativament les distintes parts d'un tot i ens donen únicament informació relativa entre les seues components (Egozcue & Pawlowsky-Glahn, 2011a) (Pawlowsky-Glahn et al., 2015). D'aquesta definició s'extreuen algunes conseqüències immediates:

- les dades composicionals apareixen en forma de vectors de dues o més components positives tot i que freqüentment se'n suprimeix una de les seues components.
- Sols els quocients entre les components aporten informació, el que exclou les dades nul·les, que no poden aportar informació relativa. Els valors nuls requereixen un tractament detallat (veure apartat 8.2).

A partir de les consideracions anteriors, es formulen els principis fonamentals següents:

1. **Invariància per escala**: Suposa que la composició no es veu afectada si, per exemple, queda expressada per tant per 1 o tant per cent. La informació continguda en ambdós casos és completament equivalent.

2. **Coherència subcomposcional:** suposa que a l'examinar un subconjunt de les parts d'una composició (una subcomposició) és necessari que els resultats de l'anàlisi no siguen contradictoris amb els obtinguts de la composició original.

3. **Escala relativa simetritzada**: cadascuna de les parts d'una subcomposició té escala relativa

4. **Invariància per permutació**: les conclusions d'un anàlisi composicional no han de dependre de l'ordenació de les parts.

Les dades composicionals, per pròpia definició, són dades en les quals les distintes parts de la composició poden adoptar únicament valors positius. Es tracta d'un subespai de la recta real de l'espai euclidià al qual anomenem Símplex. Més precisament, ens referim al Símplex de D parts (S^D) quan el nombre de parts de la composició és D.

El símplex de D parts és un subconjunt de l'espai real de D dimensions. Per a D=2 pot ser representat per un segment lineal; per a D=3 tindrem un triangle (és el conegut diagrama ternari) i per a D=4 és un tetraedre.

Per satisfer aquests principis fonamentals de l'anàlisi composicional és necessari definir una geometria del símplex de D parts. El desenvolupament dels conceptes proposats per Aitchison han dut a l'anomenada geometria d'Aitchison del símplex, de tipus euclidià, que es basa en les operacions de pertorbació i potenciació.

Tal i com hem assenyalat a la introducció de l'anàlisi composicional realitzada a l'apartat 5.6 d'aquesta tesina, d'acord amb Egozcue et al. (Egozcue & Pawlowsky-Glahn, 2011a), l'anàlisi de dades composicionals pot reduir-se als tres passos següents:

1. La transformació de les dades a coordenades tipus log-quocient;
2. L'anàlisi estadístic ("tradicional") de les coordenades anteriors com variables reals
3. La interpretació dels models obtinguts en les pròpies coordenades, tornant a expressar els resultats en termes de composicions (és a dir, fent la transformació inversa a la realitzada al primer pas).

A continuació analitzem detalladament cadascun d'aquests passos.

7.2.1 Transformació de dades a coordenades tipus log-quocient

El canvi d'unitats d'alguna o totes les parts pot considerar-se com una pertorbació. La invariància per escala de les composicions condueix de forma natural a utilitzar quocients entre les parts (passant a treballar així en termes relatius) i, una vegada considerats, aplicar logaritmes (Egozcue & Pawlowsky-Glahn, 2011b).

Existeixen distintes alternatives per realitzar aquesta transformació de coordenades: Aitchison (John Aitchison, 1982) (J. Aitchison, 1984) va introduir les transformacions *alr* (additive-log-ratio) i *clr* (centered-log-ratio) i posteriorment Pawlowsky-Glahn i Egozcue (2001) definiren la transformació *ilr* (isomètric-log-ratio). Amb aquestes transformacions, la composició inicial queda representada per un nou vector, amb totes les seues components a la recta real de l'espai euclidià, és a dir, amb valors compresos entre $-\infty$ i $+\infty$.

El fet que la transformació *ilr* conduesca a un vector de coordenades en un sistema ortogonal (cosa que les altres dos no fan) fa que les tècniques estadístiques clàssiques sobre el vector transformat puguen ser utilitzades de manera directa (Pawlowsky-Glahn & Egozcue, 2006). És per açò que utilitzarem aquesta transformació.

Cal tenir en compte, tanmateix, que la transformació *ilr* redueix la dimensió del vector transformat en una unitat respecte a la dimensió de la composició. Aquest mateix fenomen es produeix amb la transformació alr però no per a la transformació *clr.* Així, amb la transformació *ilr* passarem d'un vector de dimensió D que representa a la composició de D parts, a un altre vector de dimensió D-1 que representa la composició en coordenades transformades (totes les seues components es troben a la recta Real).

Les expressions de càlcul de les coordenades *ilr* són relativament complexes i, en general, poden no resultar fàcils d'interpretar si no es prenen certes precaucions. Per simplificar aquesta interpretació es pot escollir la base ortonormal de l'espai transformat de manera adequada, en funció del problema del que es tracte.

Una de les tècniques de construcció d'aquesta base es basa en realitzar una partició seqüencial binària (realització de balanços entre components) de forma que cada partició (o balanç) corresponga a una de les coordenades *ilr* (Egozcue & Pawlowsky-Glahn, 2011a).

Cada partició, d'un total de D-1, donarà lloc a una coordenada *ilr*, l'estructura de la qual facilita la interpretació. Així, podrem construir una matriu (matriu de contrastos) on cada fila representarà un dels balanços i a les columnes situarem les distintes parts de la composició.

En aquesta matriu, per a cada balanç s'assignaran els valors "+1", "-1" ó "0" a les distintes parts. El "0" indica que la part en qüestió no intervé en el balanç. Així, aquest queda definit per la partició dels "+1" i dels "-1" (balanç de les parts amb "+" contra les parts amb "-").

Cadascuna de les particions donarà lloc a un balanç de la forma:

$$b_j = \sqrt{\frac{r \cdot s}{r+s}} ln \frac{g_m(x_+)}{g_m(x_-)} \qquad \text{Equació (5)}$$

On $g_m(x_+)$ i $g_m(x_-)$ són les mitjanes geomètriques de les parts indicades amb signe "+" i "–" en la partició j-èsima i "r", "s" són el numero de parts amb signes "+" i "-" respectivament.

Així, la composició inicial (x1, x2, ..., x_D) donarà lloc al vector de balanços (b1, b2, ..., b_{D-1}).

7.2.2 Anàlisi estadístic tradicional de les coordenades tipus log-quocient

El vector de balanços resultat de la transformació de la composició és un vector de dimensió D-1 amb totes les seues components pertanyent a l'espai euclidià Real. Així, per a cada composició corresponent a cadascun dels mostrejos realitzats s'obtindrà un vector de balanços.

Particularitzarem l'anàlisi per a 1 país. Per a aquest tindrem una sèrie de dades formades per les N composicions associades als N mostrejos disponibles per a aquest país a la sèrie de dades de partida que estiguem analitzant (Aigua Rural/Urbana o Sanejament Rural/Urbà). Cada composició tindrà D

parts. A partir d'aquesta sèrie, se n'obtindrà una altra aplicant la transformació *ilr*. Tindrem N vectors de D-1 balanços cadascun.

A cadascun d'aquests D-1 balanços se li ajustarà un model de regressió lineal simple, de la forma descrita a l'apartat 7.1: determinarem els paràmetres del model, representarem la recta de regressió, obtindrem el coeficient de determinació i realitzarem la comprovació de les hipòtesis del model a partir de l'anàlisi gràfica de la distribució dels residus.

Amb el model lineal elaborat, es realitzarà la predicció dels valors del balanç a futur per a distints escenaris temporals.

7.2.3 Transformació inversa: recuperant el vector en les parts originals.

Si x^* representa el vector transformat (x^*=ilr(x)) predit pel model resultant de l'ajust realitzat, fent la *ilr* inversa pot tornar-se a obtenir les dades (prediccions) en les components originals:

$$x = ilr^{-1}(x) \qquad \text{Equació (6)}$$

L'expressió per obtenir la *ilr* inversa suposa utilitzar els conceptes de permutació i potenciació definits per Aitchison en el Simplex.

7.3 Comparació de resultats

A partir de les dades originals (d'un determinat tipus –aigua o sanejament- en un entorn determinat –rural o urbà-) i particularitzant per a un país en concret, tindrem un conjunt de N composicions de D parts. Per a cada part, per tant, un conjunt de N observacions. Amb aquestes dades procedirem de la següent manera:

A) Ajust d'un model lineal sobre les composicions de manera directa (mètode del JMP): Per a cadascuna de les parts de la composició realitzarem un ajust lineal dels N valors observats en funció del temps. Una vegada el model definit realitzarem la predicció de cadascuna de les parts per a distints escenaris futurs. Aquest és el model que assimilarem al del JMP.

B) Transformació, ajust del model lineal i transformació inversa (anàlisi composicional): Per a cadascuna de les parts de la composició realitzarem la transformació ilr havent definit prèviament la matriu de contrastos. Tindrem un conjunt de N vectors de D-1 balanços. Per a cada balanç tindrem N valors observats en funció del temps. Realitzarem un ajust lineal de cada balanç en funció del temps. A continuació realitzarem la valoració del model de regressió ajustat, obtindrem les rectes de regressió associades i la predicció del balanç a futur (distints escenaris temporals) per a tots ells. Seguidament realitzarem la transformació inversa dels D-1 balanços predits pels models lineals per obtenir les composicions de D parts estimades pel model.

C) Comparació:
   a. Representació gràfica de cada part: Una primera comparació de resultats es realitzarà comparant visualment l'ajust lineal realitzat a l'apartat A amb l'ajust realitzat a l'apartat B. Veurem per a cada part de la composició els resultats obtinguts per un i altre mètode.
   b. Comparació (quantitativa) dels valors estimats a futur (distints escenaris) pels dos mètodes.

8 DESENVOLUPAMENT I ANÀLISI DE RESULTATS

Com hem vist a l'apartat 6, a partir del conjunt de dades facilitades pel JMP, hem elaborat 4 sèries de dades (accés a aigua rural –WR-, accés a aigua urbana –WU-, ús d'instal·lacions de sanejament rural –SR- i de sanejament Urbà –SU-), de natura composicional.

Si agafem com a conjunt complet el total de la població (rural o urbana) d'un país, podem subdividir-lo entre la proporció dels que tenen accés a una font millorada d'aigua (per exemple) i els que no. En aquest cas el conjunt de valors pertany al subespai mostral que hem anomenat Símplex, i que serà de 2 parts. Si dividim el conjunt de la població entre aquells que tenen una font canalitzada d'aigua a la casa, aquells que utilitzen altra font millorada, aquells que s'abasteixen d'una font superficial i la resta, que s'abasteixen d'altra font no millorada, estem identificant una composició de 4 parts.

Per al cas de l'accés a les instal·lacions de sanejament pot fer-se de la mateixa forma: separar el total de la població entre la proporció que utilitza una instal·lació de sanejament millorada i els que no (utilitzen, per tant, una no millorada) o bé realitzar 4 divisions (composició de 4 parts) entre aquells que utilitzen una instal·lació de sanejament millorada individual, aquells que utilitzen una millorada tot i que compartida, aquells que defequen a l'aire lliure i la resta, que utilitza una instal·lació no millorada de sanejament.

Al nostre cas, anem a treballar amb composicions de 4 parts. L'anàlisi es realitzarà país per país de manera separada.

Així, per a cadascuna de les quatre sèries possibles (WR, WU, SR i SU), per a cada país tindrem un conjunt de N composicions de quatre components, corresponent als N mostrejos realitzats pel JMP al país en qüestió. Per a un país determinat, la composició corresponent al mostreig j-èsim serà la següent:

$$x_j=(x_1, x_2, x_3, x_4) \qquad \text{Equació (7)}$$

Donat que els x_i són parts d'una composició, comunament la seua suma és constant. Al nostre cas, cadascuna d'aquestes parts està expressada en tant per 1 pel que la suma serà la unitat.

$$\sum_{i=1}^{4} x_{i_j} = 1 \qquad \text{Equació (8)}$$

L'equació 8 ens dóna la relació entre les quatre parts de la composició per a cada observació. L'anàlisi composicional pot realitzar-se amb tres de les quatre parts o amb la composició completa. Al nostre cas considerarem aquesta última opció (composició completa), i treballarem amb el conjunt de les seues 4 parts.

Cal assenyalar que el vector x_j que representa la composició per al mostreig j-èsim és un vector de components estrictament positives de suma la unitat. La presència de valors nuls en alguna de les composicions de la sèrie de dades originals es tracta de manera específica a l'apartat 8.2.

A partir de les quatre sèries de dades de partida, la comparació dels resultats entre les dues metodologies d'anàlisi (estadística clàssica vs anàlisi composicional) s'ha realitzat a escala país, per a un conjunt de països de la regió d'Àfrica Subsahariana.

Una primera anàlisi de les dades corresponents als països de l'Àfrica Subsahariana mostren la presència de zeros (0) en algunes observacions. La presència de valors nuls en alguna de les parts de les composicions observades condiciona l'anàlisi composicional donat que aquest es basa en relacions logarítmiques, resultant impossible el seu tractament de manera directa (Martín-Fernández, Barceló-Vidal, & Pawlowsky-Glahn, 2003).

Degut a açò, a l'anàlisi comparativa realitzada que presentem a continuació, s'han exclòs aquells països amb presència de valors nuls en alguna de les observacions de la sèrie. A l'apartat 8.2 realitzarem una anàlisi específica de la presència de zeros a la sèrie de dades.

8.1 Països sense cap zero a la sèrie de dades

Els països per als quals s'ha realitzat l'anàlisi són els següents:

Aigua (rural i urbà)	Ghana	Kenya	Madagascar	Moçambic	Camerun	Rwanda			
Sanejament (rural i urbà)	Ghana	Kenya	Madagascar	Moçambic	Camerun	Rwanda	Etiòpia	Botswana	Malawi

Taula 1. Conjunt de països analitzat (sense presència de zeros a la sèrie) per a les distintes sèries de dades disponibles (WR, WU, SR i SU)

Per a cada país s'ha ajustat directament un model lineal a la sèrie de dades corresponent a cadascuna de les 4 parts de la composició. Aquest model representa el mètode del JMP.

A continuació hem realitzat l'anàlisi composicional, basada en la transformació ilr de les dades, l'ajust d'un model lineal a cada balanç i la transformació inversa per recuperar les estimacions en forma de composicions i poder-les comparar amb el mètode del JMP descrit al paràgraf anterior.

Com a exemple de l'aplicació de l'anàlisi composicional a les dades del JMP, presentem a continuació de manera detallada la metodologia particularitzada per a les dades d'accés a aigua en entorn rural (WR) a Ghana. Als annexos 1 (dades d'aigua) i 2 (dades de sanejament) es recull l'anàlisi per a la resta de països considerats.

8.1.1 Accés a l'aigua en entorn rural en Ghana

8.1.1.1 Dades de partida

La sèrie de dades disponible està formada per 17 composicions corresponents a 17 mostrejos distints (taula 2). Cada mostreig té associat també l'any en el qual es va realitzar. Cal tenir en compte que aquests mostrejos no corresponen necessàriament a distints anys, podent tenir més d'una composició per a un mateix any.

Cadascuna de les composicions és de 4 parts, on aquestes parts corresponen als distints tipus de fonts d'aigua potable que hem definit anteriorment (canalitzada a l'interior de la parcel.la (*Piped(R)*), altre tipus de font millorada (*Other_impr(R)*), font superficial (*Surface(R)*) i altra font no millorada (*Oth_Unimpr(R)*). Considerem per tant, la composició completa. La suma de les parts de la composició és la unitat. Així, tenim 17 observacions per a cadascuna de les 4 parts de la composició.

	Year	Piped(R)	Other_impr(R)	Surface (R)	Oth_Unimpr (R)
1	1988	0.01880000	0.2647923	0.6691000	0.04730764
2	1993	0.02000000	0.4485325	0.3660000	0.16546750
3	1995	0.01000000	0.4900000	0.4100000	0.09000000
4	1997	0.02092569	0.4746821	0.3933606	0.11103154
5	1998	0.03500000	0.5335241	0.3240000	0.10747588
6	1998	0.03449184	0.5107884	0.3509742	0.10374558
7	2000	0.04312823	0.4934216	0.3451769	0.11827319
8	2003	0.02300000	0.6320000	0.2750000	0.07000002
9	2003	0.01926973	0.5233891	0.3040000	0.15334114
10	2003	0.05650000	0.6120000	0.2070000	0.12449999
11	2006	0.02700000	0.6626474	0.2210000	0.08935262
12	2007	0.02016053	0.7312549	0.1782711	0.07031348
13	2008	0.02325874	0.7417413	0.1820000	0.05300000
14	2008	0.02185506	0.7239652	0.1730038	0.08117593
15	2010	0.05149614	0.6520941	0.2288316	0.06757810
16	2011	0.01894505	0.6650550	0.2150000	0.10099998
17	2013	0.02527240	0.7184756	0.1810000	0.07525198

Taula 2. Sèrie de dades d'accés a aigua a Ghana en entorn rural.

8.1.1.2 Ajust lineal

Considerem cadascuna de les 4 parts de la composició per separat i realitzem un ajust lineal de la sèrie de valors observats respecte a l'any d'observació. La recta de regressió s'ha representat per al període 1985 a 2030.

Als gràfics següents es representen les dades observades respecte a l'any d'observació i la recta de regressió corresponent a l'ajust lineal d'aquestes dades. En abscisses trobem el temps (en anys) i en ordenades el valor de la part de la composició, compresa entre 0 i 1. Cal fer atenció a l'escala de representació en ordenades, distinta segons la part de la qual es tracte.

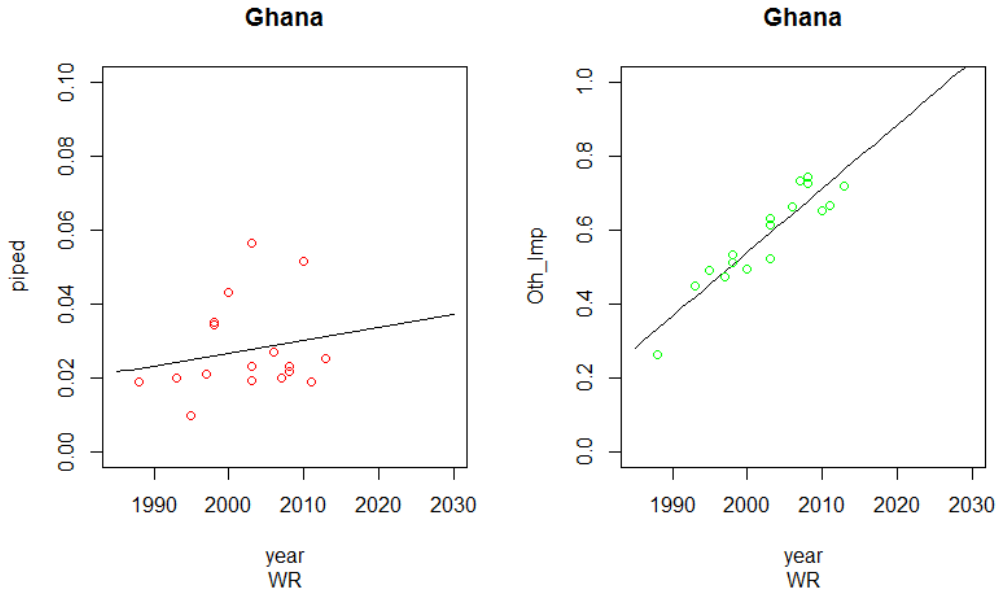


Figura 13. Ajust lineal directe dels valors observats de les parts "Piped" (gràfic de l'esquerra) i "Other Improved" (gràfic dret), de la composició de dades d'accés a una font d'aigua potable per a Ghana en entorn Rural.

A la figura 13 es mostra l'ajust per a les dues parts relacionades amb l'accés a una font millorada d'aigua potable. Podem observar que el model lineal reprodueix una tendència creixent de les dos, és a dir, amb el pas del temps més percentatge de població rural tendeix a abastar-se d'una font millorada d'aigua. Com correspon a un entorn rural en un país amb baix índex de desenvolupament, el percentatge de població abastant-se d'una font canalitzada a les seues propietats (figura 13 esquerra) és molt baix, inferior al 10% en tots els casos (compte amb l'escala del gràfic).

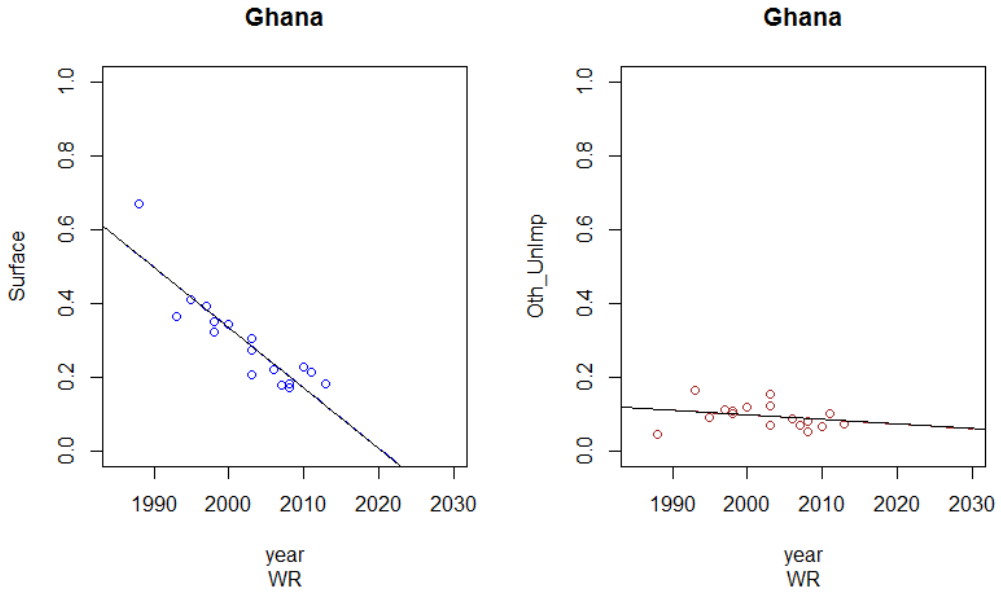


Figura 14. Ajust lineal directe dels valors observats de les parts "Surface" (gràfic de l'esquerra) i "Other Unimproved" (gràfic dret), de la composició de dades d'accés a una font d'aigua potable per a Ghana en entorn Rural.

A la figura 14 veiem la representació de les parts associades a les fonts no millorades d'aigua per al cas de Ghana en entorn Rural. El model lineal indica un decreixement en la proporció de població

abastant-se d'una font no millorada, ja siga superficial (figura 14 esquerra) o d'altre tipus (figura 14 dreta). Pot observar-se que el model lineal no té en compte la limitació dels valors a un rang determinat, com podem apreciar per al cas de l'abastament d'una font superficial (figura 14 esquerra). En efecte, tot i que els valors han d'estar compresos entre 0 i 1 (representen proporcions respecte el total), els valors estimats pel model més enllà de l'any 2020 són negatius.

Aquesta és la raó per la qual el JMP corregeix l'ajust lineal amb la introducció de dos rectes de pendent horitzontal als extrems inicial i final de la sèrie, com s'ha comentat a l'apartat 4.3.

Els valors estimats pels models lineals ajustats per a cadascuna de les 4 parts, als anys 2010, 2015, 2020, 2025 i 2030 són els següents:

Any	Piped	Oth_Impr	Surface	Other_Unimp
2010	0.03020437	0.7117741	0.171613682	0.08640781
2015	0.03192416	0.7978903	0.089974700	0.08021084
2020	0.03364396	0.8840065	0.008335717	0.07401387
2025	0.03536376	0.9701226	-0.073303266	0.06781690
2030	0.03708356	1.0562388	-0.154942249	0.06161992

Taula 3. Valors predits per a distints anys en cadascuna de les parts de la composició. Ajust lineal directe. Accés a l'aigua en Ghana en entorn Rural.

Com s'aprecia a la taula 3, tal i com hem comentat anteriorment, el model de regressió lineal en sí mateix no te en compte la natura composicional de les dades (podem apreciar a la taula 3 l'existència de valors predits fora del rang (0,1), sense cap sentit a l'hora d'interpretar-los).

8.1.1.3 Anàlisi composicional

8.1.1.3.1 Transformació *ilr* de coordenades

Cal escollir en primer lloc la matriu de contrastos que ens permetrà interpretar millor els resultats. Tal i com s'ha assenyalat a l'apartat 7.2, la construcció d'aquesta matriu es realitza seguint un procés seqüencial binari de la següent forma: el primer balanç distingirà entre l'ús de fonts millorades de les que no ho són; el segon balanç consistirà en distingir a dins dels que utilitzen fonts millorades entre els que tenen una xarxa canalitzada a l'interior de la propietat (Piped) i els que utilitzen altra font millorada distinta (Oth_Improved). Per últim, el tercer balanç es centrarà en aquells que s'abasteixen d'una font no millorada fent la distinció entre els que utilitzen una font superficial (Surface) i els que s'abasteixen d'una font no millorada de tipus distint a la superficial (Oth_Unimproved).

La matriu de contrastos és la que es mostra a la taula 4. A aquesta s'inclou un text interpretatiu de cada balanç així com els valors de *r* i *s* per al càlcul de les coordenades transformades amb la transformació *ilr*, d'acord amb l'equació 5 (veure 7.2.1).

Balanç	Piped on premises	Other Improved	Surface	Other Unimproved	Interpretació	r	s
Ilr_1 (x)	+1	+1	-1	-1	Millorades vs No millorades	2	2
Ilr_2 (x)	+1	-1	0	0	Canalitzada vs altra millorada	1	1
Ilr_3 (x)	0	0	-1	+1	Superficial vs altra no millorada	1	1

Taula 4. Matriu de contrastos

Particularitzant l'expressió de la transformació *ilr* per als balanços escollits i per a la sèrie de dades formada per les 17 composicions (taula 2) obtindrem l'expressió d'aquestes composicions en coordenades transformades dins l'espai euclidià real. Cada coordenada representa un dels balanços descrits anteriorment. A la taula 5 es mostren els valors dels balanços per a cada observació.

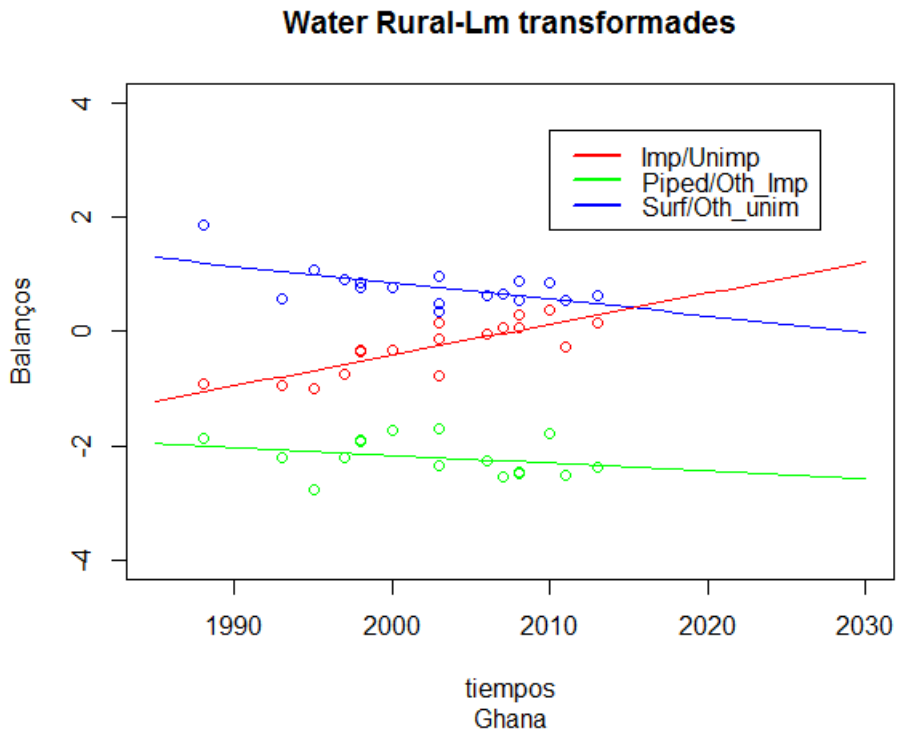
Observació	Any	BALANÇOS		
		Imp/Unimp	Piped/Oth_Imp	Surf/Oth_unim
1	1988	-0.92490138	-1.870360	1.8733110
2	1993	-0.95484736	-2.199278	0.5613428
3	1995	-1.00948817	-2.751933	1.0722196
4	1997	-0.74045907	-2.207352	0.8944282
5	1998	-0.31157881	-1.926269	0.7802761
6	1998	-0.36298810	-1.905817	0.8618014
7	2000	-0.32575622	-1.723351	0.7573538
8	2003	-0.14044155	-2.342924	0.9675170
9	2003	-0.76541606	-2.334718	0.4839174
10	2003	0.14697425	-1.684676	0.3595024
11	2006	-0.04933669	-2.263029	0.6403362
12	2007	0.08111001	-2.539245	0.6578510
13	2008	0.29069139	-2.448230	0.8723681
14	2008	0.05962186	-2.475093	0.5350640
15	2010	0.38771263	-1.795119	0.8624599
16	2011	-0.27217303	-2.516117	0.5342317
17	2013	0.14375266	-2.366983	0.6205957

Taula 5. Sèrie de dades corresponent a l'accés a l'aigua en Ghana en medi rural en coordenades transformades (transformació ilr)

8.1.1.3.2 Ajust del model lineal a les coordenades transformades

A la taula 5 tenim una sèrie de 17 observacions corresponents a cadascun dels tres balanços que hem definit. Ajustem un model de regressió lineal per a cadascun d'aquests.

Al gràfic 7 hem representat els valors observats de cadascun dels balanços i la recta de regressió del model lineal corresponent.



Gràfic 7. Models de regressió lineal per als balanços (composicions en coordenades transformades -transformació *ilr*-). Accés a l'aigua rural en Ghana.

Com podem observar, el primer balanç (millorades vs no millorades) mostra una tendència creixent i els altres dos decreixents.

El primer balanç representa el balanç entre aquells que utilitzen fonts millorades i aquells que no. Que siga creixent indica que amb el pas del temps augmenta la proporció de gent abastant-se d'una font millorada d'aigua potable i, per tant, la proporció dels que s'abasteixen d'una font no millorada disminueix.

El segon balanç serveix per comparar entre sí els dos tipus d'infraestructures millorades. Que siga decreixent suposa que la proporció dels que tenen la infraestructura canalitzada al domicili respecte als que s'abasteixen d'altra font millorada baixa. Tot i que pot semblar que aquest resultat és contradictori amb els ODM, no és així. Aquesta tendència estaria indicant un augment de la proporció d'aquells que utilitzen altra infraestructura millorada respecte als que la tenen canalitzada. Açò pot ser degut a que s'ha aconseguit pujar un primer esglaó a l'escala d'accés a l'aigua (pas de font no millorada a millorada tot i que no canalitzada) per a una gran proporció de gent, tot i que no s'està a l'esglaó superior (infraestructura canalitzada a dins de la propietat).

El tercer balanç compara els dos tipus d'usuaris de fonts d'abastament no millorades. El descens pot explicar-se pel descens d'aquells que s'abasteixen d'una font superficial respecte del total dels que utilitzen fonts no millorades. Tenint en compte que l'abastament de font superficial representa l'esglaó més baix de l'escala d'accés a l'aigua, aquesta tendència explicaria una millora de la situació tot i no arribar encara a abastar-se d'una font millorada d'aigua potable.

8.1.1.3.2.1 Recta de regressió. Coeficient de determinació

Per als tres balanços, els paràmetres del model (equació (1)) són els que es recullen a la taula 6.

Balanç	β_0	β_1
1 : (Imp/Unim)	-108.48757	0.05403899
2 : (Pip/Oth_imp)	24.61939	-0.01339209
3 : (Unimp/Oth_unimp)	58.78110	-0.02896343

Taula 6. Paràmetres del model de RLS per als tres balanços. Dades d'aigua rural en Ghana

Els resultats del model lineal per a cada balanç, obtinguts amb ajuda del software R (R Development Core Team, 2011), són els següents:

1ER BALANÇ: ACCÉS A FONTS MILLORADES vs NO MILLORADES *(IMPROVED vs UNIMPROVED)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.51794 -0.16698  0.08384  0.15469  0.39445

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.085e+02  1.937e+01  -5.600 5.07e-05 ***
timR         5.404e-02  9.675e-03   5.585 5.21e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2678 on 15 degrees of freedom
Multiple R-squared:  0.6753,      Adjusted R-squared:  0.6536
F-statistic:  31.2 on 1 and 15 DF,  p-value: 5.206e-05
```

El resultat del test F indica que no es pot rebutjar que la relació (balanç 1 en funció del temps) siga lineal. El p-valor baix del test t per a cadascun dels coeficients del model indica que tots dos són significatius.

El valor del coeficient de determinació és de 0.65 el que ens indica una relació que pot aproximar-se a la lineal (sense ser massa gran aquesta correlació)

2ON BALANÇ: FONT CANALITZADA A LA PROPIETAT vs ALTRE TIPUS MILLORAT (*PIPED vs OTHER IMPROVED)*

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.65410 -0.17630 -0.08274  0.21174  0.52029

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.61939   23.00446   1.070    0.301
timR        -0.01339    0.01149  -1.166    0.262

Residual standard error: 0.318 on 15 degrees of freedom
Multiple R-squared:  0.08307,     Adjusted R-squared:  0.02194
F-statistic: 1.359 on 1 and 15 DF,  p-value: 0.2619
```

El nivell de significació és molt baix (valor del p-valor alt), és a dir les dos variables (balanç 1 respecte del temps) no estan molt relacionades. El valor del coeficient de determinació és de 0.02194. Es tracta d'un valor baixíssim que indica que no existeix correlació entre elles. A més, els p-valors del contrast t són elevats, el que indica que els coeficients del model no són significatius suggerint que el balanç podria ser constant en el temps.

A països com Ghana en entorn rural (països amb baix índex de desenvolupament) és previsible trobar taxes molt baixes associades a la part corresponent a la font canalitzada d'aigua a l'interior de la propietat. Aquesta situació condiciona el balanç, adoptant aquest valors molt baixos, independentment del valor de la component corresponent a altres fonts millorades. Açò explicaria els resultats obtinguts on no tan sols el balanç resulta independent del temps sinó que aquest es veu tant condicionat pels valors baixos del percentatge dels "piped" que podria considerar-se constant en el temps.

3ER BALANÇ: FONT SUPERFICIAL D'AIGUA vs ALTRE TIPUS DE FONT NO MILLORADA (*SURFACE vs OTHER UNIMPROVED*)

```
Residuals:
     Min       1Q   Median       3Q      Max 
-0.49564 -0.09689 -0.04012  0.14288  0.67151 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 58.78110   20.47103   2.871   0.0117 *
timR        -0.02896    0.01022  -2.833   0.0126 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.283 on 15 degrees of freedom
Multiple R-squared:  0.3486,      Adjusted R-squared:  0.3052 
F-statistic: 8.027 on 1 and 15 DF,  p-value: 0.01259
```

El contrast F indica que existeix relació lineal i els p-valors del contrast t, tot i no ser molt baixos, indiquen que els dos coeficients del model són significatius.

El valor del coeficient de determinació és de 0.3052. Es tracta d'un valor que indica una correlació baixa entre les variables.

8.1.1.3.2.2 Comprovació d'hipòtesis

Per realitzar la comprovació de les hipòtesis del model analitzem la distribució dels residus de manera gràfica. Es presenten 4 gràfics per a cadascun dels balanços:

- Gràfic Residus-Valors Predits ("Residuals vs Fitted"), que permet jutjar la linealitat del model.
- Gràfic Quantil-Quantil ("Normal Q-Q"), que permet jutjar la hipòtesi de distribució normal dels residus
- Gràfic "Scale-Location", que permet jutjar la hipòtesi d'homoscedasticitat
- Gràfic "Residuals vs Leverage", del qual podem visualitzar la presència d'observacions crítiques, que degut a la seua importància condicionen el model.

1ER BALANÇ: ACCÉS A FONTS MILLORADES vs NO MILLORADES *(IMPROVED vs UNIMPROVED)*

La figura 15 conté els quatre gràfics d'anàlisi dels residus per al primer balanç.

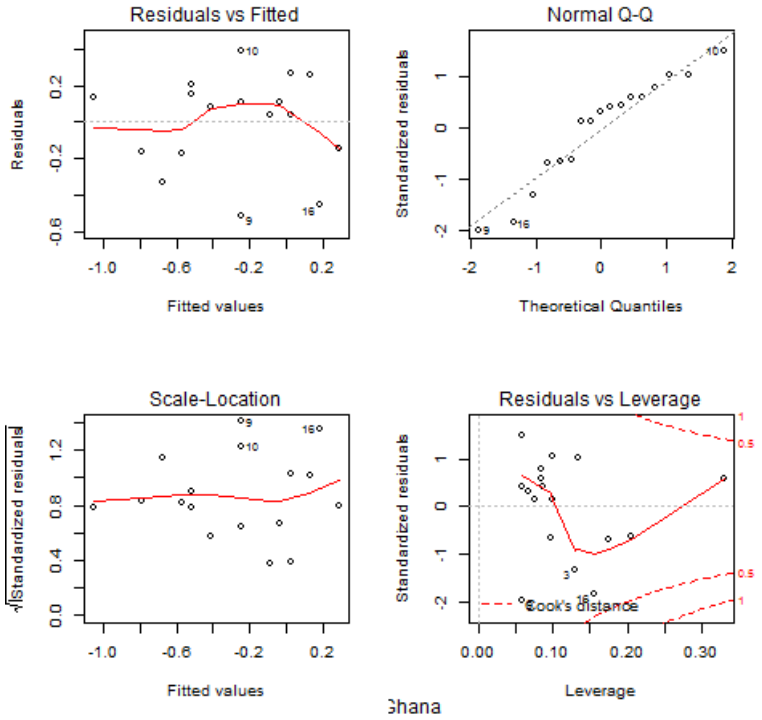


Figura 15. Comprovació de les hipòtesis del model lineal del Balanç 1 (transformació ilr, balanç font Millorada/No millorada). Dades d'aigua rural en Ghana

A partir dels gràfics anteriors podem dir que no s'aprecia una no linealitat (gràfic "Residuals vs Fitted", figura 15 dalt-esquerra) donat que els residus pareixen distribuir-se aleatòriament respecte als valors estimats.

Respecte a la hipòtesi de distribució normal dels residus, el gràfic "Normal Q-Q" (figura 15 dalt dreta) mostra que aquesta hipòtesi és discutible (els valors no es troben clarament alineats respecte a la recta de punts del gràfic). La interpretació gràfica de l'ajust resulta insuficient donat que alguns punts s'allunyen de la recta corresponent a la distribució de referència. Per confirmar-ho, caldria complementar el diagnòstic amb un contrast d'hipòtesis, com per exemple el test de Shapiro-Wilk (Shapiro & Wilk, 1965). Els resultats d'aquest test sobre els residus del model ajustat confirmen el no rebuig de la hipòtesi de normalitat de residus (p-valors alts)

```
> shapiro.test(residuals(LM1rur))

        Shapiro-Wilk normality test

data:  residuals(LM1rur)
W = 0.9341, p-value = 0.2542
```

La hipòtesi d'homoscedasticitat no es pot rebutjar ja que els punts al gràfic "Scale-Location" (figura 15 baix-esquerra) semblen distribuir-se aleatòriament de manera uniforme, tot i que entre els valors observats -1 i -0.5, la presència d'un nombre més reduït de punts no permet veure-ho amb claredat.

Per últim, no s'observen valors crítics que estiguen condicionant el model, donat que al gràfic "Residuals vs Leverage" (figura 15 baix-dreta) la distància de Cook dels valors és petita (tots els punts es troben a l'interior de les línies d'iso-distància de Cook de 0.5).

2ON BALANÇ: FONT CANALITZADA A LA PROPIETAT vs ALTRE TIPUS MILLORAT (*PIPED vs OTHER IMPROVED)*

La figura 16 conté els quatre gràfics d'anàlisi dels residus per al segon balanç.

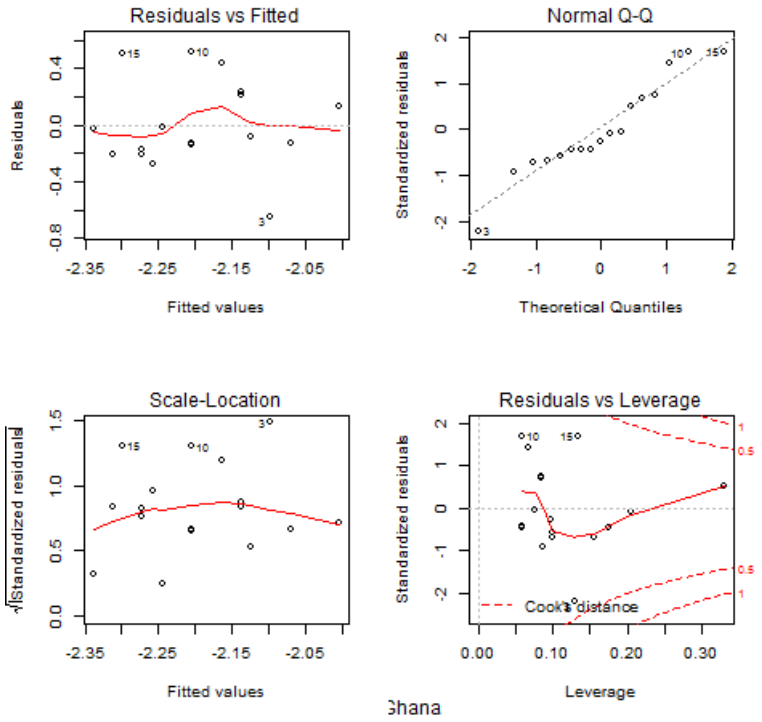


Figura 16. Test d'hipòtesis del model lineal del Balanç 2 (transformació ilr, balanç Piped/Other_Impr). Dades d'aigua rural en Ghana.

A partir dels gràfics anteriors podem dir que no s'aprecia una no linealitat (gràfic "Residuals vs Fitted", figura 16 dalt esquerra) donat que els residus pareixen distribuir-se aleatòriament respecte als valors estimats. Trobem, tanmateix, almenys tres valors que es troben "anormalment" allunyats de la mitjana (zero).

Del gràfic "Normal Q-Q" (figura 16 dalt-dreta), no podem rebutjar la hipòtesi de distribució normal dels residus. El test de Shapiro-Wilk sobre els residus ha confirmat la hipòtesi de la normalitat en la distribució d'aquests.

```
> shapiro.test(residuals(LM2rur))

        Shapiro-Wilk normality test

data:  residuals(LM2rur)
W = 0.9361, p-value = 0.2744
```

De la mateixa manera, la hipòtesi d'homoscedasticitat no pot rebutjar-se ja que els punts al gràfic "Scale-Location" (figura 16 baix-esquerra) semblen distribuir-se aleatòriament de manera uniforme.

Per últim, no s'observen valors crítics que estiguen condicionant el model, donat que al gràfic "Residuals vs Leverage" (figura 16 baix-dreta) la distància de Cook dels valors és petita (tots els punts es troben a l'interior de les línies d'iso-distància de Cook de 0.5).

3ER BALANÇ: FONT SUPERFICIAL D'AIGUA vs ALTRE TIPUS DE FONT NO MILLORADA (*SURFACE vs OTHER UNIMPROVED*)

La figura 17 conté els quatre gràfics d'anàlisi dels residus per al tercer balanç.

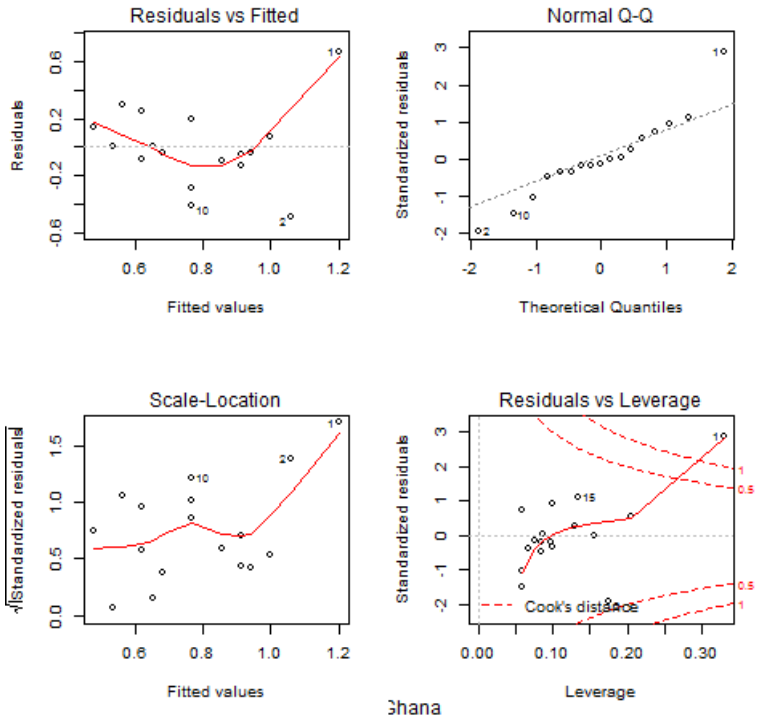


Figura 17. Test d'hipòtesis del model lineal del Balanç 3 (transformació ilr, balanç Surface/Other_UnImpr). Dades d'aigua rural en Ghana.

Un primer anàlisi d'aquests gràfics mostra la presència d'un valor anòmal, que possiblement condiciona els resultats. És el primer valor de la sèrie, etiquetat com a "1" als gràfics de la figura 17, corresponent a la composició de l'any 1988. El valor del balanç 3 per a aquesta observació és anormalment alt (taula 5), com a conseqüència d'uns valors corresponents al percentatge de gent utilitzant una font d'aigua superficial molt més alts que la mitja (0.67 per a aquesta observació front a un valor mitjà del conjunt sense aquesta observació de 0.27 –veure taula 2-).

Aquest valor podria explicar-se bé per un error en la mesura, donat que la dada és molt anterior a la definició dels Objectius del Mil·lenni i podria donar-se que el tipus de font observat no es corresponguera al tipus real; o bé que es tracte d'un valor real, tot i que anòmal, degut a una situació socioeconòmica particular[4]. Un anàlisi en profunditat de les dades d'aquest país suposaria eliminar

[4] Al cas de Ghana, la dècada dels anys '80 correspon a un període de crisi al país i una època marcada pel programa d'ajustos estructurals imposat pel Fons Monetari Internacional (FMI) com a contraprestació als préstecs realitzats per aquest. El cost social del programa d'ajust fou elevat, augmentant la fam, la mortalitat infantil i l'analfabetisme. Les migracions camp-ciutat foren una

aquest valor de la sèrie i analitzar les seues implicacions, a partir de la comparació dels resultats obtinguts amb i sense ell. En tot cas, donat que l'anàlisi que realitzem a aquesta tesina s'integra dins d'una discussió metodològica general aplicable a la totalitat de països dels quals el JMP disposa de dades, no entrarem en l'anàlisi de dades particularitzat per a aquest cas ni en la comparació de resultats entre el model amb aquesta observació i sense ella.

Així, per a la totalitat de la sèrie d'observacions, l'anàlisi gràfica dels residus mostra que existeix una dependència de l'error respecte al valor predit (gràfic de "Residuals vs Fitted", figura 17 dalt-esquerra) el que pot indicar una dependència no lineal entre les dades (balanç i temps).

Pel que fa a la hipòtesi de distribució normal dels residus, el gràfic "Normal Q-Q" (figura 17 dalt-dreta) mostra que no pot rebutjar-se aquesta hipòtesi ja que trobem un alineament dels punts al voltant de la recta. El test de Shapiro-Wilk confirma aquest fet.

```
> shapiro.test(residuals(LM3rur))

        Shapiro-Wilk normality test

data:  residuals(LM3rur)
W = 0.9574, p-value = 0.5826
```

La hipòtesi d'homoscedasticitat tampoc pot rebutjar-se ja que els punts al gràfic "Scale-Location" (figura 17 baix-esquerra) semblen distribuir-se aleatòriament de manera uniforme, tot i l'existència d'alguns valors anòmals, com el designat per "1" al gràfic, tal i com ja hem comentat. La influència d'aquest valor es confirma amb el gràfic "Residuals vs Leverage" (figura 17 baix-esquerra), presentant una distància de Cook superior a tots els altres, amb un valor major que la unitat.

8.1.1.3.2.3 Interpretació

La comprovació visual de les hipòtesis del model realitzada anteriorment condueix a afirmar que:

- El model de regressió lineal establert per al balanç 1 (Millorades vs No millorades) reprodueix raonablement bé el comportament de la variable respecte del temps. Podem assumir que aquesta es distribueix linealment. El test de Shapiro realitzat sobre els residus del model confirma que no pot rebutjar-se la hipòtesi de la seua distribució normal.
- La relació entre el balanç 2 i el temps possiblement no és de tipus lineal simple. El coeficient de determinació és molt baix, no existint una relació entre el aquest balanç i el temps. Els gràfics i els p-valors associats als coeficients del model semblen indicar que el balanç es manté constant amb el temps. Aquesta situació podria ser deguda a l'existència de valors molt baixos de la part corresponent a la font canalitzada a la propietat. Aquests valors farien que, a pesar que la part corresponent a altres fonts millorades varie en el temps, el balanç continués invariable.
- La relació entre el balanç 3 i el temps presenta molt probablement un comportament no lineal, com es dedueix de la distribució dels residus del model. Aquesta relació ve

constant, formant-se barriades pobres, sense aigua potable ni sanejament. Com a resposta, el govern inicià un programa de re-assentament en zones rurals de 12 000 persones a l'any (Hamed, 2008).

molt marcada per la influència del punt 1, corresponent a la part de la composició del mostreig realitzat al 1988. Aquesta influència es confirma per la seua distància de Cook (superior a la unitat), que ens indica que podria estar condicionant el model. Aquest valor anòmal (en relació al conjunt de la sèrie) pot donar-se per un error en la mesura (cal tenir en compte que la definició de fonts millorades i no millorades realitzada pel JMP és molt posterior) o bé per una excepcionalitat històrico-social al país. De fet, la fi de la dècada dels anys '80 a Ghana correspon a un període de forta crisi econòmica i social com a conseqüència dels plans d'ajust estructural vinculats als préstecs del Banc Mundial al país.

8.1.1.3.3 Transformació inversa.

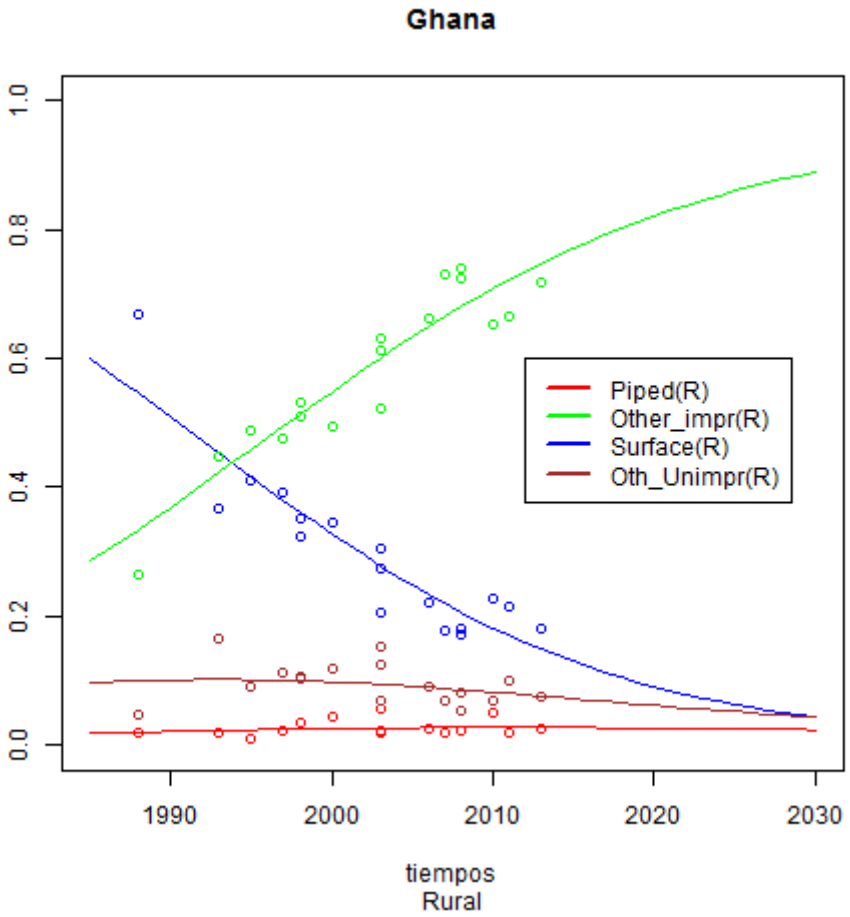
El model de regressió lineal ajustat per a cada balanç prediu els valors d'aquest en funció del temps. Hem agafat la sèrie temporal dels anys compresos entre el 1985 i el 2030, obtenint els valors any a any. El vector de balanços estimat per a cadascun dels anys de la sèrie ha estat transformat, utilitzant la transformació ilr inversa i, utilitzant la matriu de contrastos de la taula 4, s'han obtingut les composicions estimades recollides a la taula 7.

	Any	Piped(R)	Other_impr(R)	Surface(R)	Oth_Unimp(R)
1	1985	0.01775371	0.2854252	0.59986818	0.09695290
2	1986	0.01839057	0.3013169	0.58225215	0.09804038
3	1987	0.01902293	0.3176369	0.56434238	0.09899778
4	1988	0.01964823	0.3343507	0.54618285	0.09981820
5	1989	0.02026387	0.3514200	0.52782059	0.10049557
6	1990	0.02086723	0.3688028	0.50930527	0.10102472
7	1991	0.02145574	0.3864541	0.49068865	0.10140151
8	1992	0.02202684	0.4043262	0.47202405	0.10162287
9	1993	0.02257810	0.4223693	0.45336570	0.10168687
10	1994	0.02310719	0.4405319	0.43476815	0.10159278
11	1995	0.02361193	0.4587614	0.41628559	0.10134106
12	1996	0.02409031	0.4770051	0.39797121	0.10093333
13	1997	0.02454053	0.4952105	0.37987657	0.10037242
14	1998	0.02496100	0.5133257	0.36205105	0.09966223
15	1999	0.02535036	0.5313007	0.34454126	0.09880773
16	2000	0.02570749	0.5490871	0.32739060	0.09781485
17	2001	0.02603154	0.5666393	0.31063883	0.09669037
18	2002	0.02632190	0.5839145	0.29432179	0.09544185
19	2003	0.02657821	0.6008732	0.27847111	0.09407744
20	2004	0.02680034	0.6174797	0.26311409	0.09260583
21	2005	0.02698842	0.6337019	0.24827358	0.09103610
22	2006	0.02714277	0.6495116	0.23396803	0.08937757
23	2007	0.02726392	0.6648848	0.22021155	0.08763973
24	2008	0.02735261	0.6798013	0.20701402	0.08583208
25	2009	0.02740970	0.6942449	0.19438135	0.08396405
26	2010	0.02743621	0.7082033	0.18231563	0.08204490
27	2011	0.02743331	0.7216675	0.17081549	0.08008365
28	2012	0.02740223	0.7346324	0.15987637	0.07808898
29	2013	0.02734430	0.7470957	0.14949081	0.07606921
30	2014	0.02726091	0.7590580	0.13964885	0.07403220
31	2015	0.02715349	0.7705229	0.13033829	0.07198537
32	2016	0.02702350	0.7814958	0.12154504	0.06993561
33	2017	0.02687241	0.7919848	0.11325347	0.06788934
34	2018	0.02670167	0.8019993	0.10544664	0.06585241
35	2019	0.02651275	0.8115504	0.09810663	0.06383021
36	2020	0.02630706	0.8206506	0.09121479	0.06182755
37	2021	0.02608600	0.8293133	0.08475195	0.05984879

38	2022	0.02585090	0.8375527	0.07869865	0.05789778
39	2023	0.02560308	0.8453837	0.07303535	0.05597790
40	2024	0.02534378	0.8528216	0.06774255	0.05409211
41	2025	0.02507418	0.8598819	0.06280096	0.05224294
42	2026	0.02479543	0.8665804	0.05819162	0.05043252
43	2027	0.02450859	0.8729328	0.05389599	0.04866263
44	2028	0.02421468	0.8789545	0.04989605	0.04693472
45	2029	0.02391465	0.8846611	0.04617435	0.04524991
46	2030	0.02360939	0.8900675	0.04271406	0.04360906

Taula 7. Composicions estimades en funció de l'any a partir de l'anàlisi composicional. Dades d'Aigua en Ghana en entorn rural

Si realitzem la representació gràfica dels valors anteriors i superposem els valors observats de les composicions originals tenim un gràfic com el següent (gràfic 8):



Gràfic 8. Representació del model de regressió composicional i de les composicions observades. Dades d'accés a l'aigua en Ghana en entorn rural.

L'anàlisi del gràfic 8 permet constatar les tendències que s'intuïen a l'ajust dels balanços:

- L'evolució en quan a l'accés a una font d'aigua mostra que la proporció amb accés a fonts millorades augmenta enormement a costa de reduir la proporció de gent que usa fonts no millorades d'aigua
- Aquest augment es deu fonamentalment a l'increment de la proporció en l'ús de fonts millorades de tipus no canalitzat (pous d'aigua millorats, per exemple). La

cobertura pel que fa a xarxes d'abastament domiciliar és escassa (com bé correspon a un entorn rural de l'Àfrica subsahariana) i la tendència es manté gairebé constant al llarg del temps.

- Els valors corresponents a proporció de gent utilitzant fonts alimentades per xarxes canalitzades a l'interior de la propietat són molt baixos durant tot el període (rarament supera el 10% respecte al total) el que confirmaria que també el segon dels balanços (canalitzada vs altre tipus millorat) fóra constant al llarg del temps.
- La reducció en l'ús de fonts no millorades es deu sobretot a que cada vegada menys proporció de gent s'abasteix d'una font superficial.

Pel que fa a l'anàlisi del gràfic en la seua totalitat cal dir que:

- El model no és lineal pel que fa a l'ajust de les composicions: trobem línies corbes.
- Donat que estem veient les composicions en el seu conjunt, en tot moment la suma entre elles és 1 pel que els valors de les parts sempre estan compresos entre 0 i 1. No necessitem, per tant, forçar el model per imposar valors dins del rang possible, com fa el mètode del JMP i altres mètodes alternatius (veure apartat 5).

8.1.1.4 Comparació de resultats: Model RLS vs Model composicional

8.1.1.4.1 Comparació visual: anàlisi gràfica

A la figura 18 hem representat, component a component, les dades observades i dels ajustos del model de regressió lineal simple (RLS) i del model sorgit de l'anàlisi composicional.

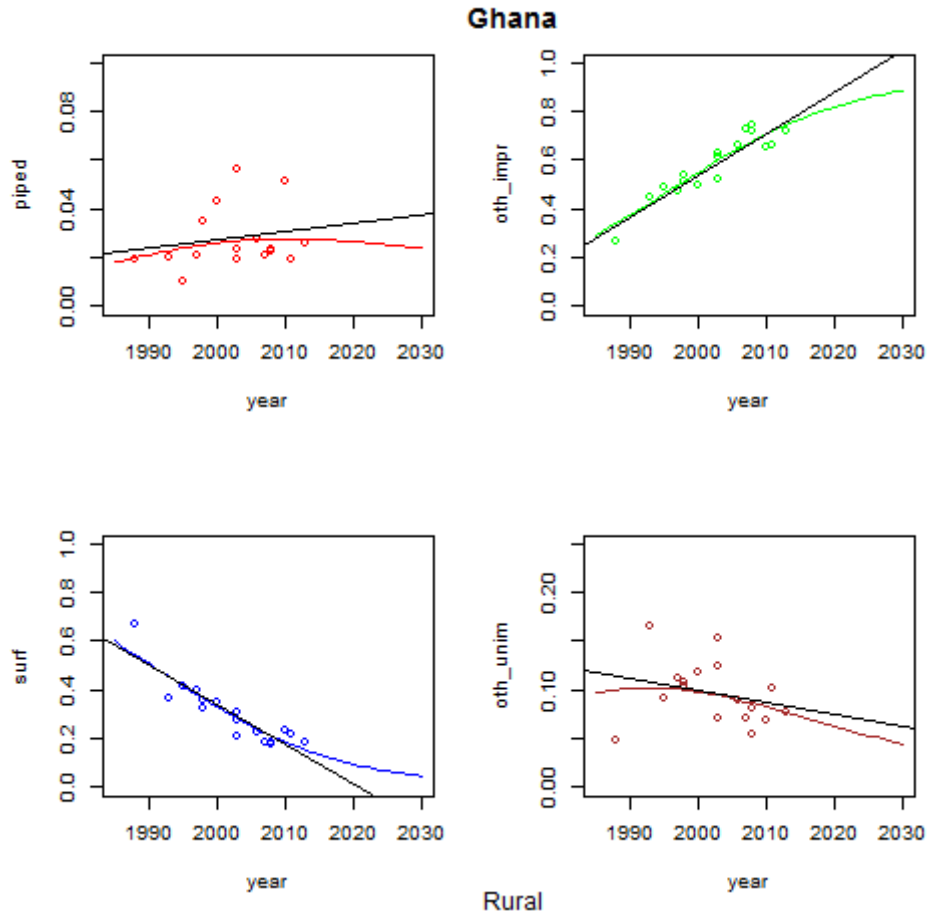


Figura 18. Model composicional (en color) i ajust lineal per a cada part de la composició. Accés a aigua en entorn rural a Ghana.

Als gràfics de la figura anterior (figura 18) es mostren:

- en colors els punts corresponents als valors observats de cadascuna de les parts de la composició per a l'any al qual correspon l'observació.
- Les corbes de color representen els models de regressió de l'anàlisi composicional per a cadascuna de les parts de la composició
- Les rectes en negre són les rectes de regressió del model lineal

Podem observar que:

- El model composicional està representat per una corba, el que possibilita la presència de màxims i mínims (relatius o absoluts) que mostren una variació en la tendència de les dades. Aquesta pot correspondre a un canvi de ritme en el creixement o decreixement (cas de les parts "other improved" –figura 18, gràfic dalt a la dreta- o "surface" –figura 18, gràfic baix a l'esquerra-) o a un canvi de tendència creixent/decreixent o viceversa (cas de les altres dos components). El model de regressió lineal simple no pot reproduir aquests tipus de canvis de tendències, donat que la pendent és única.

- La natura composicional de les dades està plenament considerada pel model composicional. Així, els models tenen en compte el comportament conjunt (suma de les components constant) per a qualsevol any en el qual es realitze l'estimació. La pendent de les corbes és variable per ajustar-se al rang de valors possibles de cadascuna de les parts. Contràriament, el model de regressió lineal simple creix o decreix indefinidament, adoptant valors impossibles (fora del rang entre 0 i 1) a partir d'un horitzó temporal determinat.

8.1.1.4.2 Comparació quantitativa: valors predits a distints escenaris temporals

A la taula 8 es recullen els valors predits pels dos models (lineal –LM- i composicional –CoDa-) per a cadascuna de les parts en distints escenaris temporals. Aquests escenaris corresponen a la sèrie d'anys entre 2010 i 2030, cada 5 anys. Les diferències absolutes entre ells per a cadascuna de les parts i relatives respecte al valor estimat per un model o altre es recullen a la taula 9.

Com pot apreciar-se a aquesta (taula 9), les diferències entre aquestes prediccions són creixents (en valor absolut) a mesura que avancem en la projecció de les estimacions.

Tal i com hem assenyalat en altres apartats, les prediccions donades pel model de regressió lineal per a alguns dels escenaris no són extrapolables a la realitat. És el cas d'aquelles parts que per a determinats escenaris temporals estan fora del rang (0,1) (valors de la part "Surface" als anys 2025 i 2030 i valor de la part "Other Improved" al 2030, a l'exemple de la taula 9).

Any	MODEL LINEAL (LM)				MODEL COMPOSICIONAL (CoDa)			
	Piped	Oth_Impr	Surface	Other_Unimp	Piped	Oth_Impr	Surface	Other_Unimp
2010	0.03020437	0.7117741	0.171613682	0.08640781	0.02743621	0.7082033	0.18231563	0.08204490
2015	0.03192416	0.7978903	0.089974700	0.08021084	0.02715349	0.7705229	0.13033829	0.07198537
2020	0.03364396	0.8840065	0.008335717	0.07401387	0.02630706	0.8206506	0.09121479	0.06182755
2025	0.03536376	0.9701226	-0.073303266	0.06781690	0.02507418	0.8598819	0.06280096	0.05224294
2030	0.03708356	1.0562388	-0.154942249	0.06161992	0.02360939	0.8900675	0.04271406	0.04360906

Taula 8. Valors predits per a distints anys en cadascuna de les parts de la composició. Ajust lineal directe.
Accés a l'aigua en Ghana en entorn rural.

Any	Piped			Other Improved			Surface			Other Unimproved		
	dif=CoDa-LM	Dif/CoDa (%)	Dif/LM (%)	dif=CoDa-LM	Dif/CoDa (%)	Dif/LM (%)	dif	Dif/CoDa (%)	Dif/LM (%)	dif=CoDa-LM	Dif/CoDa (%)	Dif/LM (%)
2010	-0.00276816	-10.1%	-9.2%	-0.0035708	-0.5%	-0.5%	0.01070195	5.9%	6.2%	-0.00436291	-5.3%	-5.0%
2015	-0.00477067	-17.6%	-14.9%	-0.0273674	-3.6%	-3.4%	0.04036359	31.0%	44.9%	-0.00822547	-11.4%	-10.3%
2020	-0.0073369	-27.9%	-21.8%	-0.0633559	-7.7%	-7.2%	0.08287907	90.9%	994.3%	-0.01218632	-19.7%	-16.5%
2025	-0.01028958	-41.0%	-29.1%	-0.1102407	-12.8%	-11.4%	0.13610423	216.7%	-185.7%	-0.01557396	-29.8%	-23.0%
2030	-0.01347417	-57.1%	-36.3%	-0.1661713	-18.7%	-15.7%	0.19765631	462.7%	-127.6%	-0.01801086	-41.3%	-29.2%

Taula 9. Diferències entre els valors predits pels models d'ajust composicional i lineal. Diferències absolutes i relatives respecte als valors predits pels dos models.
Accés a l'aigua en Ghana en entorn rural.

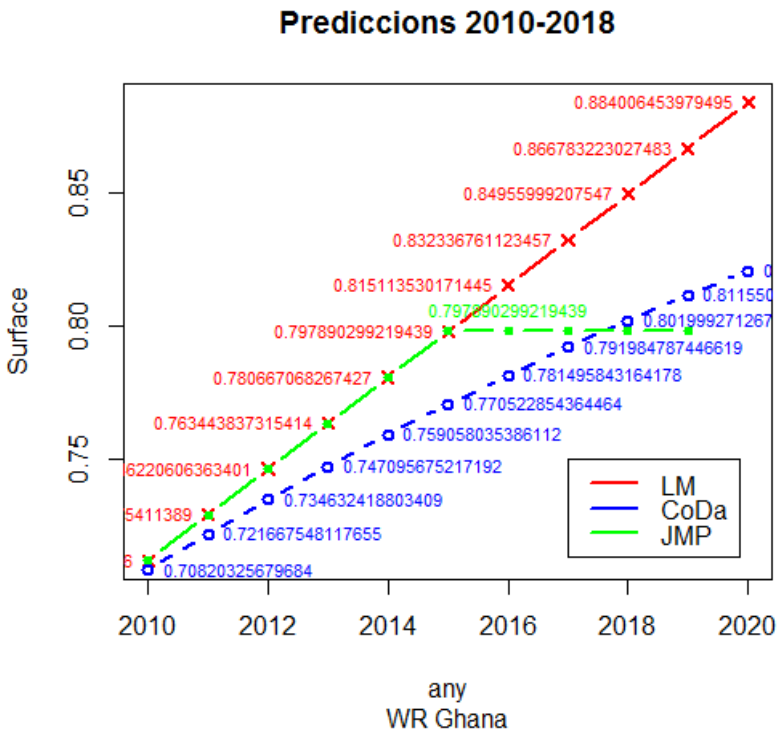
8.1.1.5 Un apunt addicional. MRLS vs JMP

L'existència de valors predits fora del rang de valors possibles fa que el model de regressió lineal simple no siga aplicable de manera generalitzada per a qualsevol escenari temporal. El JMP, conscient d'aquest fet, evita realitzar estimacions a nivell de país a més de sis anys a partir de l'últim dels valors disponibles a la sèrie d'observacions (veure 4.3). A l'exemple de les dades d'aigua en entorn rural a Ghana, l'última observació correspon a l'any 2013 (taula 2), el que suposa que el JMP realitze la predicció amb el model de regressió lineal simple fins l'any 2015 i, a partir d'aquest, mantinga constants els valors de la predicció durant 4 anys (fins al 2019). Més enllà, el JMP no proveeix estimacions.

A la taula 10 hem recollit els valors predits pels tres models dels quals estem parlant: el model de regressió lineal simple (MRLS), el model del JMP i el model composicional, per a la sèrie d'anys entre el 2010 i 2020. Aquests valors s'han representat al gràfic 9, que permet apreciar clarament la diferència entre ells.

Any	ALTRES MILLORADES (OTHER IMPROVED)		
	LM	JMP (LM modificat)	CoDa
2010	0.7117741	0.7117741	0.70820332
2011	0.7289974	0.7289974	0.72166753
2012	0.7462206	0.7462206	0.73463244
2013	0.7634438	0.7634438	0.74709575
2014	0.7806671	0.7806671	0.75905806
2015	0.7978903	0.7978903	0.77052297
2016	0.8151135	0.7978903	0.78149588
2017	0.8323368	0.7978903	0.79198489
2018	0.8495600	0.7978903	0.8019993
2019	0.8667832	0.7978903	0.8115504
2020	0.8840065	NA	0.8206506

Taula 10. Prediccions de la component Other Improved per als models de RLS, del JMP i composicional (CoDa). Sèrie temporal 2010-2020. Accés a l'aigua en entorn rural a Ghana



Gràfic 9. Prediccions de la proporció de gent abastant-se d'una font millorada (no canalitzada a la propietat) d'aigua, en entorn rural a Ghana, pels models de RLS, del JMP i composicional (CoDa) a l'interval 2010-2020.

El mètode del JMP per a una component presenta dos problemes principals:

- No permet la estandardització de la metodologia, donat que ha de limitar cas per cas les prediccions a futur, forçant el model a adoptar una pendent nul.la durant un període de temps.
- No proporciona valors més enllà de 6 anys després de l'última dada disponible (excepte si la proporció és inferior al 5% o superior al 95% que s'adopta indefinidament) el que fa que el monitoratge de l'avanç respecte a un objectiu a llarg plaç (com suposava la meta 7c dels ODM) siga impossible de realitzar.
  No cal oblidar que el monitoratge anual s'ha realitzat des de l'any 2000 i la meta fa referència a l'any 2015. Com veiem, aquest quinze anys de diferència disten molt dels sis anys des de l'última dada en els que el JMP realitza prediccions.

A més, com que cada part de la composició està relacionada entre sí, la manipulació "manual" feta pel JMP hauria d'alterar els valors de les altres parts, cosa que no necessàriament es té en compte no conservant-se així la condició de suma constant entre les parts.

L'anàlisi composicional, per contra, ressol aquests problemes: representa una metodologia que pot estandarditzar-se ja que sempre manté els valors predits al rang dels valors possibles; proporciona estimacions a futur per a qualsevol escenari temporal i; té en compte en tot moment la composició completa, pel que manté la propietat de suma constant de les seues parts.

8.2 Sèries de dades amb presència de zeros

8.2.1 Tipus de zeros

En l'anàlisi composicional es distingeix entre dos tipus de zeros: zeros "essencials" (*essential zeros)* i zeros arrodonits (*rounded zeros)* (Martín-Fernández et al., 2003).

Els zeros essencials són observacions que efectivament són zero. Aquests tipus de zeros indiquen que alguna de les components no existeix. Si estem analitzant, per exemple, com es distribueix la despesa familiar mensual, un valor nul de la despesa en tabac (per exemple) per a una família en concret seria un tipus de zero essencial (la família no gasta en tabac). La presència d'aquest tipus de zeros indica que s'està treballant amb més d'una població. Així, una manera de tractar aquests zeros és subdividir la mostra entre aquella població d'individus que tenen aquesta component nul.la i els que no. A l'exemple, es tractaria de dividir la mostra de famílies entre les que gasten en tabac i les que no. Aquesta "solució" no és aplicable quan el que tenim són zeros arrodonits.

Aquests tipus de zeros són valors que realment no són zero sinó que són valors tan petits que no els ha pogut detectar l'aparell de mesura, o bé en un tipus de mostreig com el del JMP on es conten proporcions de famílies respecte el total, que la limitació a un nombre determinat de decimals fa que els valors més petits que un determinat llindar (fixat a partir del nombre de decimals considerat) siguen arrodonits a zero. Queda clar que considerant les composicions de 4 parts referents a l'accés a una font d'aigua potable i a l'accés a un tipus d'instal·lació de sanejament que hem vist als apartats anteriors, la sèrie de dades no pot tenir cap zero essencial (l'origen de l'aigua és un tipus de font d'entre les 4 parts de la composició; i la gent defeca a un dels 4 tipus d'instal·lacions).

Per tal de tractar el problema dels zeros arrodonits, sembla raonable reemplaçar aquests zeros per un valor molt petit, de forma que permeta continuar amb l'aplicació de les relacions logarítmiques subjacents en l'anàlisi composicional (Martín-Fernández et al., 2003). Tanmateix, l'elecció d'aquest valor de reemplaçament pot condicionar els resultats de l'anàlisi (Martín-Fernández et al., 2003; Martín-Fernández, Hron, Templ, Filzmoser, & Palarea-Albaladejo, 2012), tal i com podrem veure a continuació.

8.2.2 Tractament

El procediment seguit és la substitució dels zeros observats per un valor suficientment petit que permeta aplicar l'anàlisi composicional. Cal tenir en compte que la natura composicional de les dades fa que totes entre elles estiguen interrelacionades (són parts d'un tot) i que l'augment d'una de les components (la substitució del zero per un valor, tot i que siga petit) ens obligarà a disminuir les altres ja que la suma ha de ser constant: en el nostre cas, 1. Malgrat tot, donat que en el tractament composicional les components s'analitzen conjuntament, l'efecte de la modificació serà molt menor que en els mètodes clàssics.

Existeixen distintes tècniques que permeten realitzar aquesta substitució mantenint la suma constant. Nosaltres utilitzarem l'estratègia de substitució multiplicativa (*Multiplicative Replacement Strategy)*, de la manera en que es descriu en Martín-Fernández et al., 2003.

Si tenim una composició $(x_1, x_2, ..., x_n)$ on hi ha **Z** zeros, aquesta estratègia de substitució suposa passar a la composició $(r_1, r_2, ..., r_n)$ de forma que:

$$r_j = \begin{cases} \delta_j & si\ x_j = 0, \\ \left(1 - \frac{\sum_{k|x_k=0} \delta_k}{c}\right) \cdot x_j & si\ x_j > 0 \end{cases} \qquad \text{Equació (9)}$$

On δ_j és el valor de substitució que escollim i c és la constant de la suma de les components, al nostre cas 1.

Particularitzant aquesta expressió per al nostre cas, amb 4 components d'anàlisi i amb la presència d'una única de les components de valor zero, l'expressió anterior és la següent:

$$r_j = \begin{cases} \delta_j & si\ x_j = 0, \\ (1 - \delta_j\ ) \cdot x_j & si\ x_j > 0 \end{cases} \qquad \text{Equació (10)}$$

La clau està ara en escollir el δ_j adequat. Per veure la sensibilitat dels resultats front a distints valors de δ_j i la importància d'aquest, hem analitzat el comportament per a dos valors:

- $\delta_j = 10^{-3}$
- $\delta_j = 10^{-7}$

8.2.3 Anàlisi de resultats. Sensibilitat

Per poder jutjar la sensibilitat dels resultats front al valor de substitució anem a centrar-nos en 2 països: Burkina Faso per a les dades d'aigua i Benín per a les dades de Sanejament.

El procediment ha estat el següent:

- Elecció del valor de substitució δ_j ;
- Modificació de la sèrie de dades observades per al país, modificant les composicions amb algun valor nul segons l'equació 10 per al valor de substitució escollit;
- Anàlisi composicional de la sèrie de dades modificada:
  - Transformació *ilr* de coordenades (amb la base ortonormal definida per la matriu de contrastos de la taula 4);
  - Ajust de model lineal per als tres balanços;
  - Transformació inversa per recuperar les prediccions en forma de composicions;
  - Representació gràfica del model i dels valors observats per a cadascuna de les quatre parts.

8.2.3.1 Burkina Faso

8.2.3.1.1 Entorn Rural

8.2.3.1.1.1 Sèrie de Dades originals (WR)

A la taula 11 es mostra la sèrie de dades disponibles corresponents a l'accés a l'aigua en entorn rural a Burkina Faso. S'observa la presència de tres valors nuls (ombrejats en gris a la taula) en tres de les composicions observades, tots ells en relació a la part "Piped", que representa la proporció de gent amb aigua canalitzada dins de la seua propietat. En aquest entorn és lògic pensar que el percentatge de gent amb infraestructura d'aquest tipus serà molt baix, tant baix que no supera el llindar mínim de decimals considerats, tractant-se, per tant, d'un zero arrodonit.

Any	Piped	Other_imp	Surface	Oth_unimp
1993	0.0032	0.38593965	0.048	0.56286035
1994	0.002	0.45958627	0.111	0.42741373
1996	0.0018	0.37220001	0.071	0.55499999
1996	0.00045	0.43255	0.112	0.455
1998	0.001	0.56899999	0.124	0.30600001
1999	0.0005	0.59232023	0.06	0.34717977
2003	0.0011318	0.63747195	0.05402101	0.30737525
2003	0	0.52144444	0.16	0.31855556
2003	0.0034	0.6277	0.0015	0.3674
2003	0.001	0.64299999	0.053	0.30300001
2005	0	0.71200001	0.067	0.22099999
2006	0.001	0.71699999	0.039	0.24300001
2006	0.00125	0.59274998	0.072	0.33400002
2007	0.002	0.69399998	0.031	0.27300002
2009	0.00167205	0.63347749	0.08559517	0.27925529
2010	0	0.71200001	0.079	0.20899999
2010	0.001	0.67400001	0.025	0.29999999

Taula 11. Observacions corresponents a l'accés a aigua en entorn Rural a Burkina Faso. En ombrejat gris els zeros de la sèrie.

8.2.3.1.1.2 Sèrie de Dades modificades: valor de substitució 10^-3

Realitzant la substitució de l'equació 10, per a un valor de substitució de 10^{-3}, tenim la sèrie de dades de la taula 12.

Any	Piped	Other_imp	Surface	Oth_unimp
1993	0.0032	0.38593965	0.048	0.56286035
1994	0.002	0.45958627	0.111	0.42741373
1996	0.0018	0.37220001	0.071	0.55499999
1996	0.00045	0.43255	0.112	0.455
1998	0.001	0.56899999	0.124	0.30600001
1999	0.0005	0.59232023	0.06	0.34717977
2003	0.0011318	0.63747195	0.05402101	0.30737525
2003	0.001	0.520923	0.15984	0.318237
2003	0.0034	0.6277	0.0015	0.3674
2003	0.001	0.64299999	0.053	0.30300001
2005	0.001	0.71128801	0.066933	0.22077899
2006	0.001	0.71699999	0.039	0.24300001
2006	0.00125	0.59274998	0.072	0.33400002
2007	0.002	0.69399998	0.031	0.27300002
2009	0.00167205	0.63347749	0.08559517	0.27925529
2010	0.001	0.71128801	0.078921	0.20879099
2010	0.001	0.67400001	0.025	0.29999999

Taula 12. Observacions modificades amb valor de substitució 10^{-3}. Dades d'accés a l'aigua a Burkina Faso en entorn rural. En fons gris la posició dels zeros de la sèrie original i en font blava les composicions modificades

8.2.3.1.1.3 Sèrie de Dades modificades: valor de substitució 10^-7

Realitzant la substitució de l'equació 10, en aquest cas per a un valor de substitució de 10^{-7}, tenim la sèrie de dades de la taula 13.

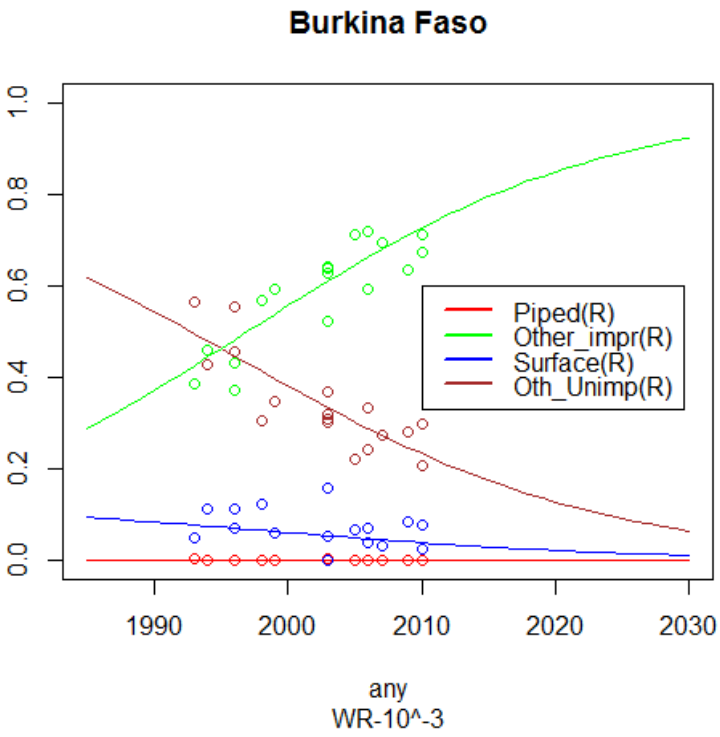
Any	Piped	Other_imp	Surface	Oth_unimp
1993	0.0032	0.38593965	0.048	0.56286035
1994	0.002	0.45958627	0.111	0.42741373
1996	0.0018	0.37220001	0.071	0.55499999
1996	0.00045	0.43255	0.112	0.455
1998	0.001	0.56899999	0.124	0.30600001
1999	0.0005	0.59232023	0.06	0.34717977
2003	0.0011318	0.63747195	0.05402101	0.30737525
2003	0.0000001	0.52144439	0.15999998	0.31855553
2003	0.0034	0.6277	0.0015	0.3674
2003	0.001	0.64299999	0.053	0.30300001
2005	0.0000001	0.71199994	0.067	0.22099997
2006	0.001	0.71699999	0.039	0.24300001
2006	0.00125	0.59274998	0.072	0.33400002
2007	0.002	0.69399998	0.031	0.27300002
2009	0.00167205	0.63347749	0.08559517	0.27925529
2010	0.0000001	0.71199994	0.079	0.20899997
2010	0.001	0.67400001	0.025	0.29999999

Taula 13. Observacions modificades amb valor de substitució 10^{-7}. Dades d'accés a l'aigua a Burkina Faso en entorn rural. En fons gris la posició dels zeros de la sèrie original i en font blava les composicions modificades

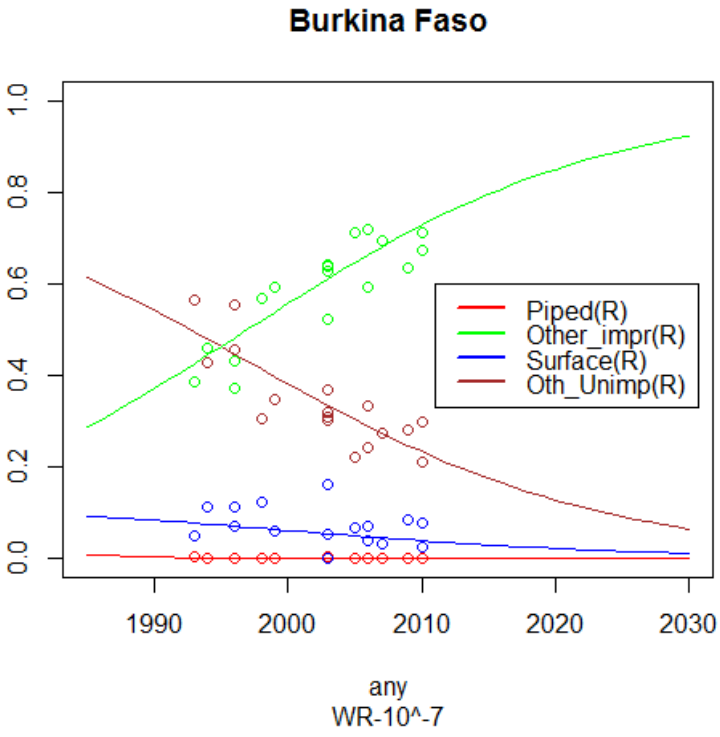
8.2.3.1.1.4 Model composicional. Resultats

Amb la sèrie de dades de les taules 12 i 13 hem realitzat l'anàlisi composicional de la mateixa manera que s'ha descrit a 8.1.1.3 (transformació *ilr*, ajust d'un model lineal a les coordenades transformades i transformació inversa).

Els resultats obtinguts es mostren als gràfics 10 i 11. Aquests mostren el model ajustat per a les distintes parts de la composició, per a les observacions de la taula 12 (valor de substitució 10^{-3}) i de la taula 13 (valor de substitució 10^{-7}). Els punts representats als gràfic corresponen a les observacions de les taules 12 i 13.



Gràfic 10. Model composicional per a sèrie de dades modificades (valor substitució 10^-3) d'accés a l'aigua en entorn rural a Burkina Faso.



Gràfic 11. Model composicional per a sèrie de dades modificades (valor substitució 10^-7) d'accés a l'aigua en entorn rural a Burkina Faso.

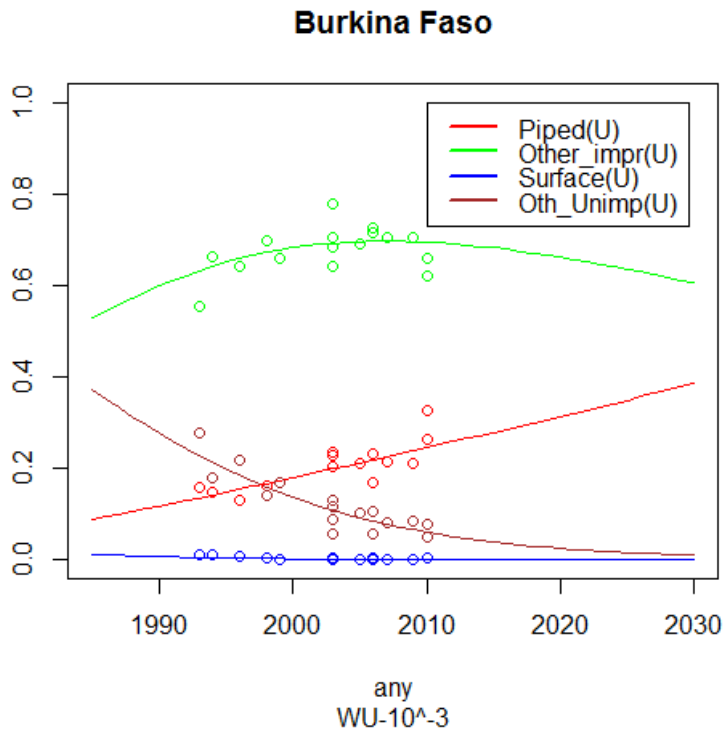
Com es pot apreciar, en aquest cas, els resultats obtinguts són molt semblants per a un valor de substitució del zero de 10^{-3} i 10^{-7}.

8.2.3.1.2 Entorn Urbà

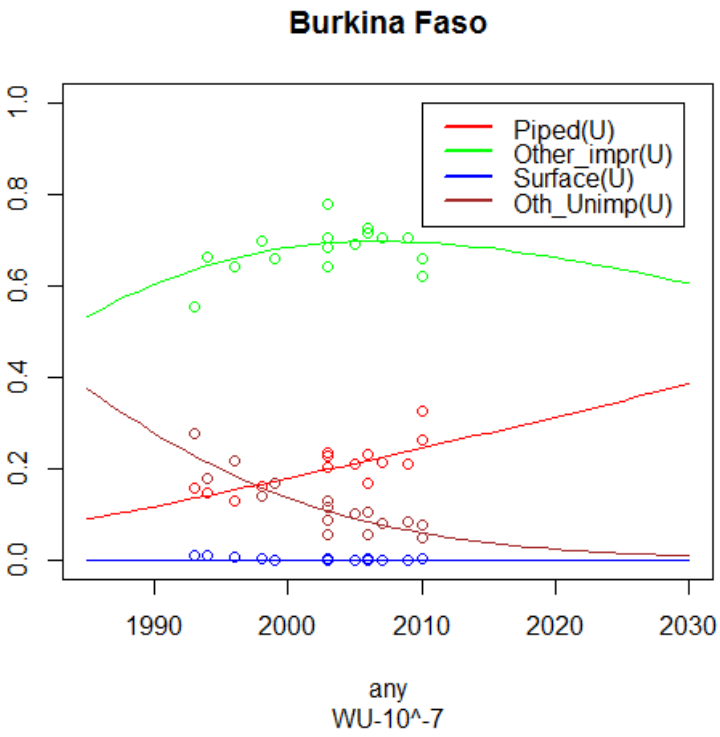
Per les dades d'accés a aigua en entorn urbà el procediment és el mateix: de la sèrie de dades disponibles per al país en entorn urbà se n'obté altra modificada, resultat de substituir la composició amb la presència d'algun zero a alguna de les parts, per altra d'acord amb l'equació 10.

De la mateixa manera que hem fet per a les dades corresponents a l'entorn rural, comparem els resultats obtinguts per a dos valors distints de substitució del zero: 10^{-3} i 10^{-7}. La sèrie de dades modificada (una per a cadascun dels dos valors de substitució) constituirà el conjunt de valors observats en cada cas. Es realitza la transformació *ilr* de les dades, l'ajust a cada part d'un model lineal en coordenades transformades (balanços) i finalment la transformació inversa del model a forma composicional i la seua representació gràfica.

Els gràfics 12 i 13 mostren els resultats obtinguts per a un i altre valor de la substitució respectivament. A cadascun d'ells trobem representat el model d'ajust de cada part resultant (línia de color) i les observacions de la sèrie (punts de color).



Gràfic 12. Model composicional per a sèrie de dades modificades (valor substitució 10^-3) d'accés a l'aigua en entorn urbà a Burkina Faso



Gràfic 13. Model composicional per a sèrie de dades modificades (valor substitució 10^-7) d'accés a l'aigua en entorn urbà a Burkina Faso

Comparant ambdós gràfics podem apreciar que el resultat és el mateix que hem vist anteriorment: les diferències entre els models d'ajust composicional per a cadascuna de les parts corresponents a un i a altre valor de substitució, no són apreciables a simple vista.

8.2.3.2 Benín

Al cas de Benín ens centrem en les dades de sanejament. Tenim quatre composicions representant l'accés a instal·lacions de sanejament. Aquestes poden ser de tipus millorat i ús individual, de tipus millorat tot i que d'ús compartit, defecació a l'aire lliure o altre tipus d'instal·lació no millorada.

La proporció mitjana d'instal·lacions millorades compartides respecte el total per al país és de 0.576978833.

8.2.3.2.1 Entorn rural

8.2.3.2.1.1 Dades originals (SR)

La sèrie de dades de sanejament observades a Benín en entorn rural són les que es mostren a la taula 14. A la sèrie, en la composició corresponent a l'any 2011 apareix un zero, a la part d'ús d'altres tipus d'instal·lacions no millorades ("Other_Unimproved").

Any	Open Defecation (OD)	Other Unimproved	Improved	Impr-Shared
1992	0.92363757	0.0366458	0.01680097	0.02291565
1996	0.91300005	0.03699995	0.02115106	0.02884894
2001	0.84699994	0.05825006	0.04008126	0.05466875
2002	0.875	0.04393671	0.03429149	0.0467718
2003	0.83599997	0.06500003	0.0418791	0.0571209
2003	0.83199996	0.04200004	0.05330067	0.07269933
2006	0.81432652	0.06534684	0.05090072	0.06942592
2009	0.76999998	0.10213891	0.05408796	0.07377315
2011	0.80300003	0	0.08333517	0.11366483
2012	0.77370566	0.09947643	0.05364666	0.07317125

Taula 14. Observacions corresponents al sanejament en entorn Rural a Benín. En fons gris els zeros de la sèrie.

8.2.3.2.1.2 Sèrie de dades modificades: substitució 10^{-3}

Si substituïm la composició corresponent al 2011 per una altra donada per l'equació 10 on el valor de substitució siga de 10^{-3} obtenim una nova sèrie, que es mostra a la taula 15.

Any	Open Defecation (OD)	Other Unimproved	Improved	Impr-Shared
1992	0.92363757	0.0366458	0.01680097	0.02291565
1996	0.91300005	0.03699995	0.02115106	0.02884894
2001	0.84699994	0.05825006	0.04008126	0.05466875
2002	0.875	0.04393671	0.03429149	0.0467718
2003	0.83599997	0.06500003	0.0418791	0.0571209
2003	0.83199996	0.04200004	0.05330067	0.07269933
2006	0.81432652	0.06534684	0.05090072	0.06942592
2009	0.76999998	0.10213891	0.05408796	0.07377315
2011	0.80219703	0.001	0.08325184	0.11355117
2012	0.77370566	0.09947643	0.05364666	0.07317125

Taula 15. Observacions modificades amb valor de substitució 10^{-3}. Dades de sanejament a Benín en entorn rural (en fons gris la posició dels zeros de la sèrie original i en font blava les composicions modificades)

8.2.3.2.1.3 Sèrie de dades modificades: substitució 10^{-7}

Per a un valor de δ_j de 10^{-7}i utilitzant l'expressió de l'equació 10, la sèrie de dades modificada és la de la taula 16.

Any	Open Defecation (OD)	Other Unimproved	Improved	Impr-Shared
1992	0.92363757	0.0366458	0.01680097	0.02291565
1996	0.91300005	0.03699995	0.02115106	0.02884894
2001	0.84699994	0.05825006	0.04008126	0.05466875
2002	0.875	0.04393671	0.03429149	0.0467718
2003	0.83599997	0.06500003	0.0418791	0.0571209
2003	0.83199996	0.04200004	0.05330067	0.07269933
2006	0.81432652	0.06534684	0.05090072	0.06942592
2009	0.76999998	0.10213891	0.05408796	0.07377315
2011	0.80299995	0.0000001	0.08333516	0.11366482
2012	0.77370566	0.09947643	0.05364666	0.07317125

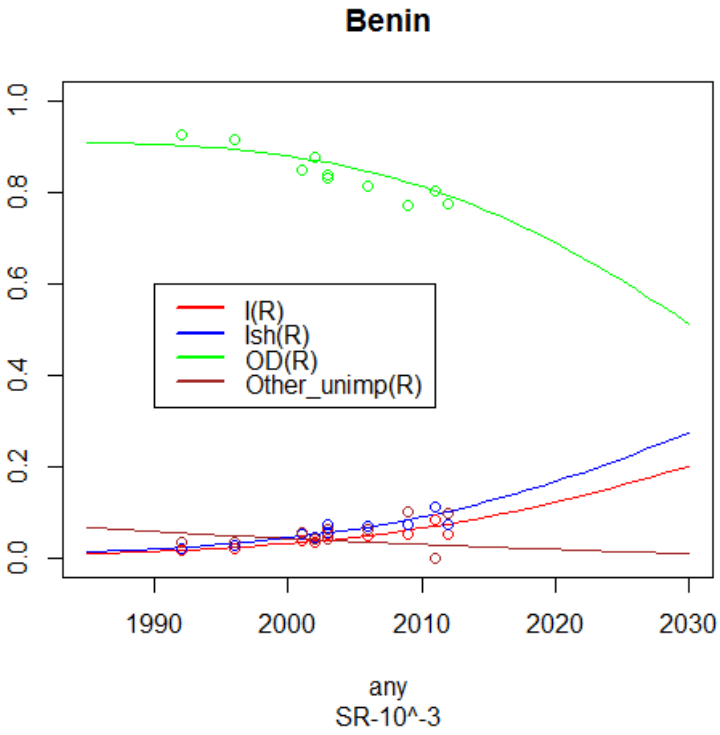
Taula 16. Observacions modificades amb valor de substitució 10^{-7}. Dades de sanejament a Benín en entorn rural (en fons gris la posició dels zeros de la sèrie original i en font blava les composicions modificades)

8.2.3.2.1.4 Model composicional. Resultats.

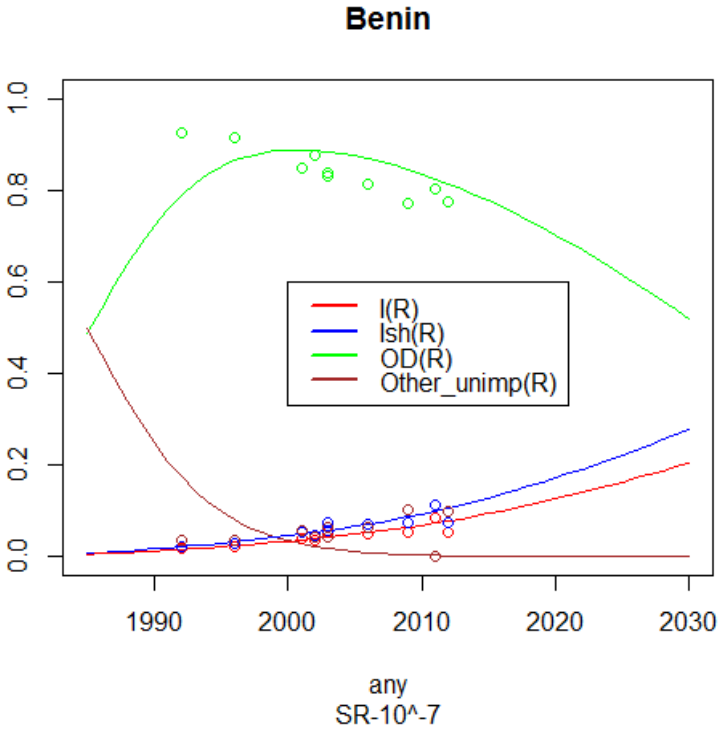
Amb la sèrie de dades de les taules 15 i 16 hem realitzat l'anàlisi composicional de la mateixa manera que s'ha descrit a 8.1.1.3 (transformació *ilr*, ajust d'un model lineal a les coordenades transformades i transformació inversa) i s'han representat aquests models d'estimació de la composició al llarg del temps. Aquesta representació la trobem als gràfics 14 i 15. Juntament amb els valors predits pel model (línies contínues als gràfics) es representen també els valors observats (punts al gràfic), que corresponen als valors de les taules 15 i 16 respectivament.

Als resultats corresponents a l'entorn Rural pot apreciar-se que són les parts corresponents a l'ús d'instal·lacions no millorades (ja siga la defecació a l'aire lliure o altre tipus de pràctica considerada com no millorada) les que més es veuen afectades pel valor de substitució escollit. En efecte, com podem apreciar als gràfics, la distribució de la proporció de gent que defeca a l'aire lliure decreix suaument i de manera contínua si la substitució és de 10^-3 (gràfic 14) mentre que creix en el període 1990-2000 per decréixer posteriorment, amb una curvatura més pronunciada, si el valor de substitució és de 10^-7 (gràfic 15). Aquest efecte el veiem igualment a la part corresponent a altres instal·lacions no millorades de sanejament. Trobem un decreixement suau i continu en un cas (substitució de 10^-3, gràfic 14) i de manera molt pronunciada durant el període 1990-2000 i més suaument en la resta, per a l'altre cas (10^-7, gràfic 15).

Com podem apreciar comparant els gràfics, els resultats obtinguts per a un valor de substitució i per a l'altre difereixen considerablement. Veiem com aquests resultats estan clarament condicionats pel valor de substitució escollit.



Gràfic 14. Model composicional per a dades de sanejament a Benín en entorn rural. Cas de dades amb zeros i valor de substitució: 10^-3



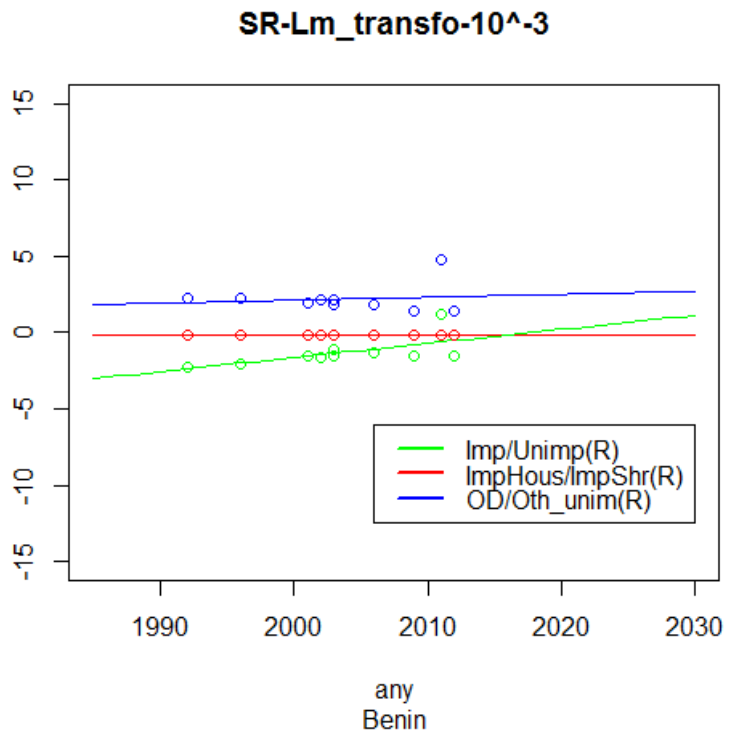
Gràfic 15. Model composicional per a dades de sanejament a Benín en entorn rural. Cas de dades amb zeros i valor de substitució: 10^-7

La interpretabilitat d'aquests resultats (el per què es produeixen) no és immediata. Tenint en compte que el model composicional té el seu origen a un model lineal en les coordenades transformades (que representen els balanços entre les seues parts) podem trobar una explicació més clara.

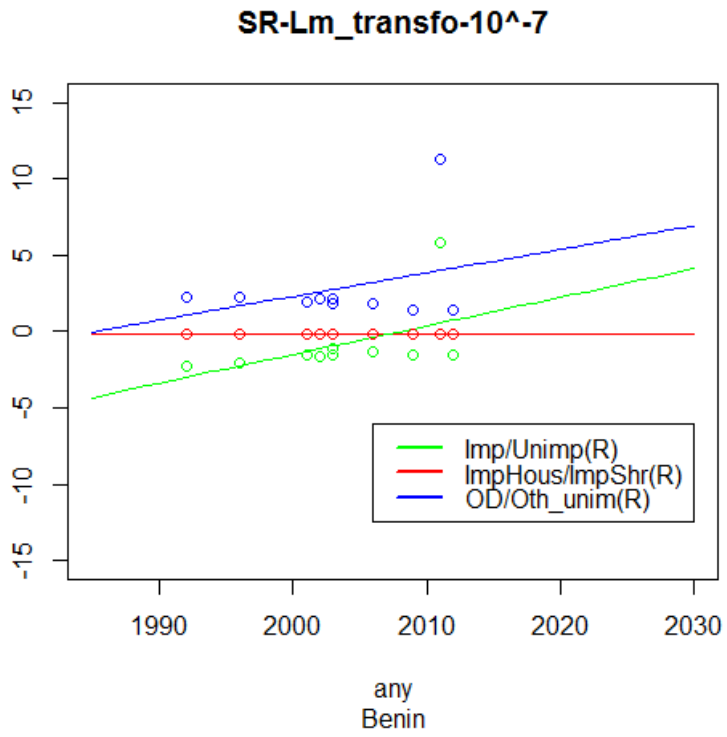
Als gràfics 16 i 17 s'han representat els models de regressió lineal dels balanços (en coordenades transformades doncs) per a les dades de sanejament rural en Benín per als dos valors de substitució.

El valor nul de la sèrie de dades original corresponia a la part d'utilització d'altres instal·lacions no millorades de sanejament, per a l'any 2011.

Si ens fixem en els balanços, trobem que les observacions corresponents al 2011 presenten menor variabilitat respecte a la recta de regressió corresponent al cas del valor de substitució de 10^{-3} (gràfic 16) que al de 10^{-7} (gràfic 17). Aquest comportament és lògic donat que a menor valor de la part "Other-Unimproved", els balanços "Open Defecation/Other_UnImproved" i "Improved/Unimproved" seran més grans. I ho són en tanta mesura que impliquen una notable variació de la pendent de les rectes de regressió associades (gràfic 17), el que suposa valors dels balanços estimats més alts a partir del 2011 i més baixos als anys anteriors. Aquesta modificació en la pendent de la recta en coordenades transformades es tradueix en un canvi de curvatura en el model composicional (gràfics 14 i 15).



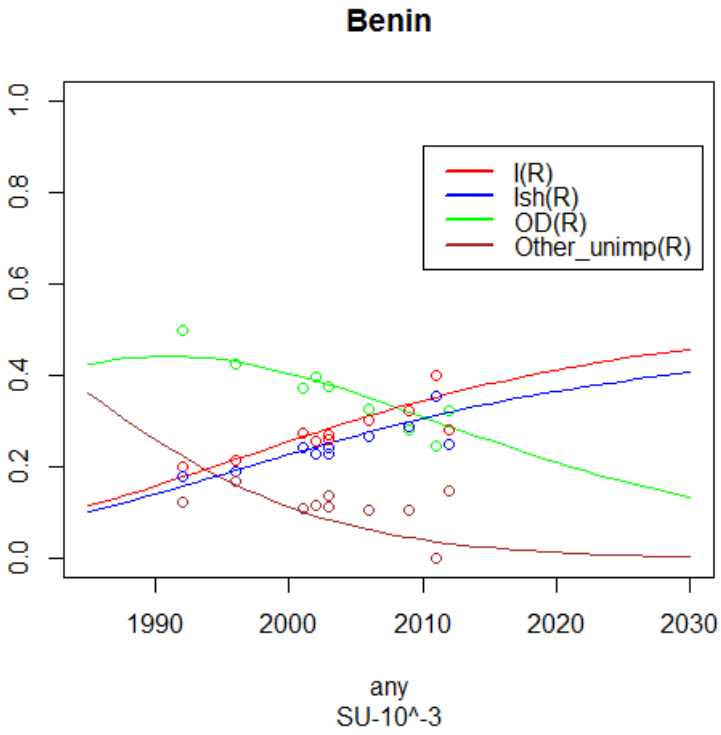
Gràfic 16. Ajust de les rectes de regressió en coordenades transformades (balanços) per a sanejament rural en Benín. Valor de substitució de 10^{-3}.



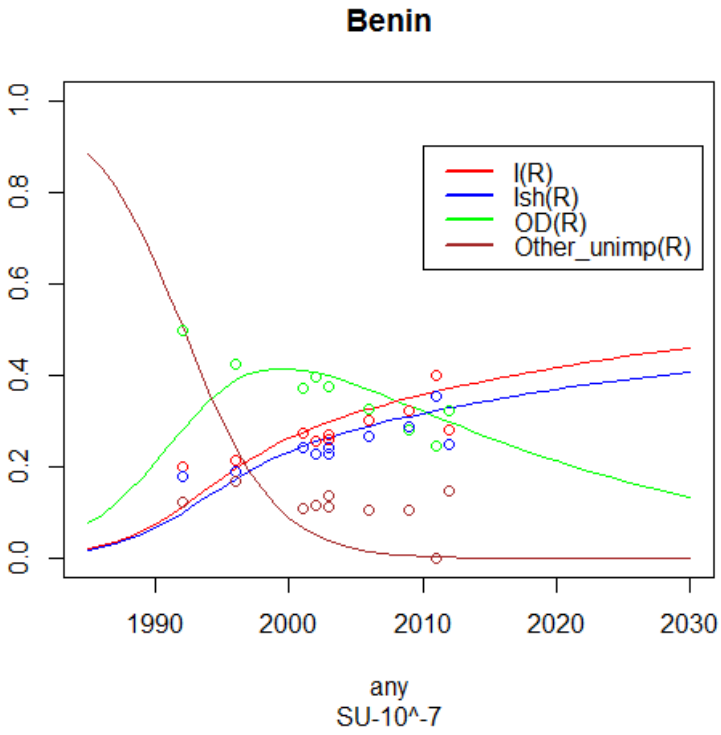
Gràfic 17. Ajust de les rectes de regressió en coordenades transformades (balanços) per a sanejament rural en Benín. Valor de substitució de 10^{-7}.

8.2.3.2.2 Entorn urbà

Als gràfics 18 i 19 hem representat el resultat del model composicional per a les 4 parts de la composició corresponent a l'accés a instal·lacions de sanejament, per a Benín en entorn urbà. De la mateixa manera que a l'apartat anterior, a més del model per a cada component, a cada gràfic es representen els punts corresponents a les observacions de la sèrie (modificada, per reemplaçar el valor zero). De la mateixa manera que passava per a l'entorn rural, podem apreciar que la influència d'escollir un o altre valor de substitució és enorme. El valor de 10^-7 (gràfic 19) corresponent a la part "Other Unimproved" per a l'any 2011 força el model a adaptar-se a ell i, donat que totes les parts de la composició estan interrelacionades, condiciona aquestes. En particular és la part corresponent a la defecació a l'aire lliure, per a la predicció en els anys anteriors al 2011, la que es veu afectada. Aquest resultat és lògic donat que ambdues parts representen una subcomposició del total, que és la utilització d'instal·lacions (o pràctiques) no millorades de sanejament.

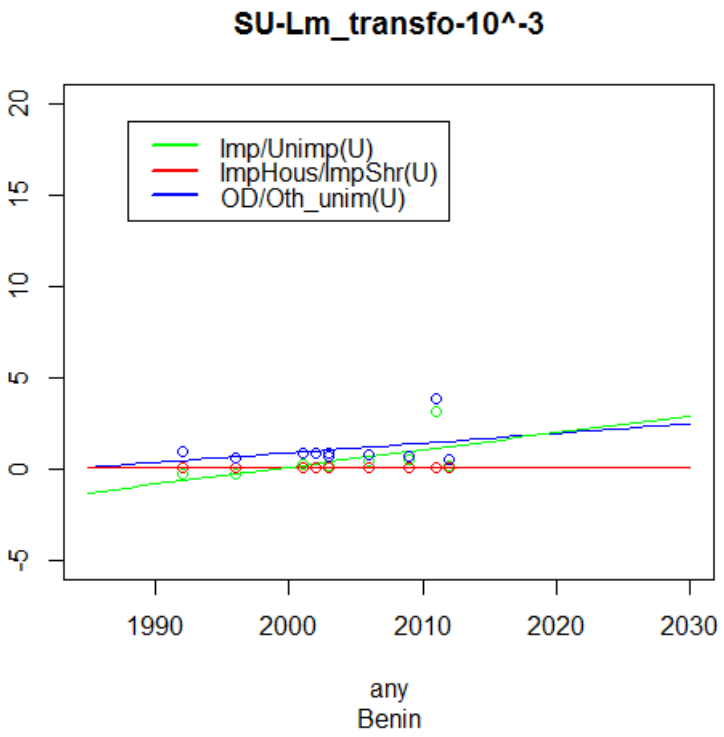


Gràfic 18. Model composicional per a dades de sanejament a Benín en entorn urbà. Cas de dades amb zeros i valor de substitució: 10^{-3}

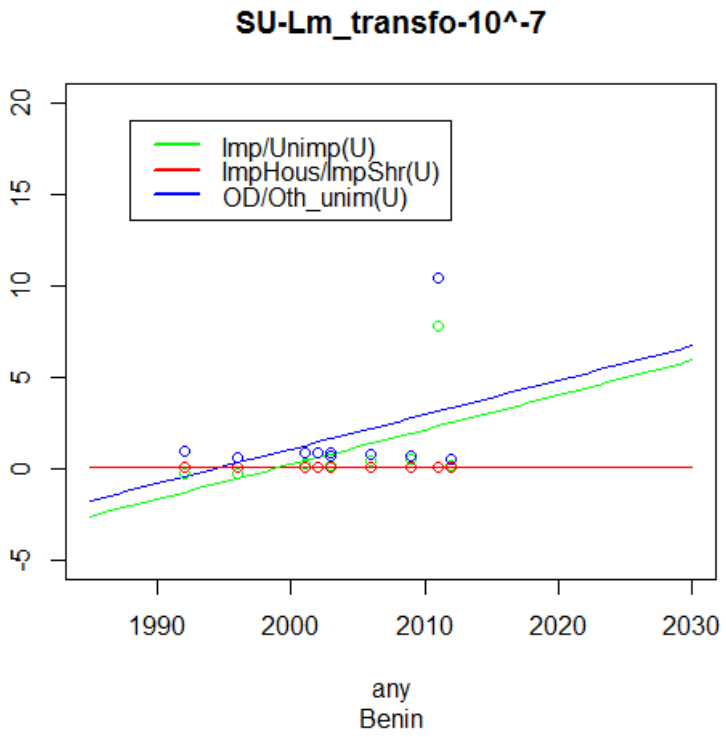


Gràfic 19. Model composicional per a dades de sanejament a Benín en entorn urbà. Cas de dades amb zeros i valor de substitució: 10^{-7}

L'anàlisi visual del model de regressió lineal realitzat sobre els balanços (en coordenades transformades) permet veure millor la influència dels valors de substitució adoptats. Als gràfics 20 i 21 poden apreciar-se els tres valors corresponent als balanços on la part nul.la en la sèrie original ha estat substituïda (any 2011). Dels tres, la posició dels dos punts corresponents al balanç "Improved/Unimproved" (punt verd) i "Open Defecation/Other Unimproved" (punt blau), és la que presenta majors diferències entre un i altre gràfic. Intuïtivament veiem que conforme aquests punts adopten valors més alts, estan forçant un augment de la pendent de la recta de regressió associada. Al gràfic 20 (valor de substitució de 10^{-3}) aquests pendents són més suaus que al gràfic 21 (substitució per 10^{-7}).



Gràfic 20. Ajust de les rectes de regressió en coordenades transformades (balanços) per a sanejament urbà en Benín. Valor de substitució de 10^{-3}



Gràfic 21. Ajust de les rectes de regressió en coordenades transformades (balanços) per a sanejament urbà en Benín. Valor de substitució de 10^{-7}

Aquest canvi de pendent del model lineal d'ajust dels balanços (gràfic 21) és el responsable del canvi brusc de curvatura en el model d'ajust de les composicions associades (gràfic 19).

9 QÜESTIONS OBERTES

A) QUANTIFICACIÓ DE LA BONDAT DE L'AJUST I COMPARACIÓ QUANTITATIVA D'AQUEST

A l'apartat 8.1 hem analitzat de manera detallada l'ajust dels models de regressió lineal i el derivat de l'anàlisi composicional, particularitzat per a les dades d'accés a aigua de la subpoblació rural d'un país determinat (Ghana, en aquest cas). Els resultats d'ambdós models s'ha comparat de manera qualitativa a partir de la representació gràfica de l'evolució de la predicció de les distintes parts de la composició (accés a aigua en entorn rural a Ghana) en el temps. A aquests gràfics s'han representat així mateix els valors de les composicions realment observades. Quantitativament hem pogut comparar els valors predits pels models per a distints escenaris temporals futurs.

A més de la comparació qualitativa en quan a l'ajust dels models respecte a les dades originals seria convenient poder quantificar l'ajust d'aquests amb algun criteri de tipus estadístic que permetera comptabilitzar en quina mesura, per a la sèrie de dades disponibles, el model composicional és millor, similar o pitjor que el model de regressió lineal simple.

Existeixen criteris estadístics de comparació de models d'ajust com el criteri d'informació d'Akaike (criteri AIC) o el criteri d'informació baiesià (BIC). Aquests criteris permeten obtenir un índex associat a un model i a la sèrie de dades original. El model amb índex inferior representa el millor ajust a la sèrie de dades (Anderson & Burnham, 2002).

Malauradament aquests criteris no són aplicables al nostre cas per a la comparació dels dos models ja que aquests no han estat ajustats per a la mateixa sèrie de dades. Mentre que el model de regressió lineal utilitza les dades observades per a la determinació de l'ajust per mínims quadrats, de la manera que hem descrit en 7.1, al cas del model composicional, l'ajust l'hem realitzat en coordenades transformades (balanços). Les prediccions del model han estat realitzades, així mateix, en aquestes coordenades i, posteriorment, transformades mitjançant la transformació ilr inversa per recuperar les composicions.

B) SENSIBILITAT ALS ZEROS. VALOR DE SUBSTITUCIÓ

La metodologia de l'anàlisi composicional que hem presentat a l'apartat 7.2 exigeix treballar en dades estrictament positives, que formen una composició de suma constant. Les distintes parts de la composició no poden ser nul·les donat que al seu tractament es produeix una transformació de coordenades utilitzant log quocients (o log-contrastos) com a base del procés (apartat 7.2).

Tanmateix, a la sèrie de dades originals podem tenir, com hem vist a l'apartat 8.2, valors nuls. Els valors nuls podem classificar-los en dos tipus segons la seua natura: zeros essencials o zeros arrodonits. Al nostre cas, únicament la presència d'aquests últims és possible. La "solució" passa per reconvertir-los a un valor no nul reemplaçant-los per un altre valor, utilitzant alguna tècnica que permeta respectar la condició de suma constant de la composició. Nosaltres hem utilitzat l'estratègia de substitució multiplicativa (veure 8) aplicada a dos casos particulars (dades d'aigua en entorn urbà i rural a Burkina Faso (apartat 8.2.3.1) i dades de sanejament en entorn rural i urbà a Benín 8.2.3.2). Siga quina siga la tècnica, l'elecció del valor de substitució del zero (o zeros) pot condicionar els resultats obtinguts (com hem vist al cas de Benín).

Seria interesant conèixer si existeix un valor de substitució comú al conjunt dels països (per exemple d'una mateixa regió), a partir del qual obtenim un model ajustat composicional que s'ajusta a les dades disponibles. D'aquesta forma podria sistematitzar-se la metodologia de substitució dels valors nuls a la sèrie de dades, què és el que interessa al cas del JMP.

Per determinar-ho caldria fer un anàlisi de sensibilitat complet (agafant múltiples valors de substitució del zero) país per país dins d'una mateixa regió i determinar el valor de substitució més adequat a cadascun. Una vegada obtinguts els valors de substitució més adequats per país caldria analitzar si existeix una relació entre els associats a tots els països d'una mateixa regió.

C) ALTERNATIVES AL MRLS EN COORDENADES TRANSFORMADES

El model composicional que hem vist és el resultat d'ajustar un model de regressió lineal simple als balanços o coordenades transformades, realitzar les previsions en aquestes coordenades i recuperar posteriorment les composicions predites mitjançant una transformació inversa.

Alternativament, caldria explorar la possibilitat que altre model distint del model de regressió lineal simple ajustés millor a la sèrie de dades en coordenades transformades (balanços).

L'elecció del model més acurat per a la sèrie de dades disponibles (sèrie de dades en coordenades transformades) podria realitzar-se, ara sí, amb les tècniques estadístiques disponibles (donat la transformació ens situa a l'espai real euclidià) com per exemple el criteri d'informació d'Akaike (AIC) o el BIC. En aquest podríem comparar els models entre sí ja que tots ells haurien estat obtinguts a partir de la mateixa sèrie de dades (balanços).

10 CONCLUSIONS

El *Joint Monitoring Program* (JMP) de Nacions Unides per a abastament d'aigua i sanejament ha estat l'organisme encarregat de monitorar el progrés cap a la consecució de la meta 7c dels Objectius del Mil·lenni, que estableix la necessitat de reduir a la meitat per a l'any 2015 la proporció de persones sense accés a una font millorada d'aigua per beure i a una instal·lació bàsica de sanejament. En preparació de l'escenari post ODM, el JMP llançà en 2014 una discussió al voltant de la metodologia utilitzada per a l'estimació dels distints indicadors.

Les dades amb les quals treballa el JMP representen proporcions de llars respecte al total de cada país amb accés a un tipus determinat de font d'aigua o d'instal·lació de sanejament. Són dades de tipus composicional sobre les quals el JMP ha vingut aplicant un model basat en el model de regressió lineal simple per al seu tractament i la realització de prediccions a futur.

Com hem vist al llarg d'aquesta tesina, l'aplicació de les ferramentes d'estadística clàssica directament sobre dades composicionals condueix a interpretacions que podrien resultar errònies. En conseqüència, les interpretacions derivades de l'anàlisi de dades i del model de regressió aplicat sobre aquestes pel JMP han de ser realitzades amb molta cura. Les propostes alternatives al mètode del JMP basades en l'aplicació d'altres models de regressió de manera directa sobre les dades, sense tenir en compte la seua natura composicional, condueixen al mateix tipus d'errors.

L'aplicació del model de regressió lineal per a realitzar prediccions a futur condueix en molts casos a l'obtenció de resultats impossibles (fora del rang de valors possibles) a partir de certs escenaris temporals i no manté la condició de suma constant entre les distintes parts de la composició.

El JMP, conscient de la limitació del model de regressió lineal simple (MRLS) pel que fa a la predicció de valors fora de rang, corregeix el MRLS limitant la predicció de valors del model únicament a dos anys a partir de l'últim valor de la sèrie. Passats aquests dos anys, en el cas general, manté el valor predit com a constant durant quatre anys més i, a partir d'ací és incapaç de predir cap valor. Les prediccions per país no s'allarguen mai més de sis anys (en el cas general) des de l'última dada. Tenint en compte que els ODM sorgeixen l'any 2000 i que la fita s'estableix per al 2015, açò suposaria no tenir cap predicció sobre l'escenari final fins l'any 2009 (6 anys enrere).

A pesar d'açò, el JMP ha realitzat estimacions del tipus "a aquest ritme, no s'alcançarà la meta 7c en una determinada regió per a les dades d'aigua (o sanejament) fins X anys després del 2015" (PNUD, 2006). Aquest tipus d'afirmacions no es sostenen en cap altra ferramenta que en l'extrapolació lineal de les dades globals que, com hem assenyalat, pot conduir a resultats erronis donada la natura composicional de les dades.

El tipus de dades del que disposem precisen d'una metodologia alternativa, adaptada a la seua natura composicional, fàcilment sistematitzable i que permeta realitzar prediccions a futur siga quin siga l'horitzó temporal, garantint la condició de suma constant entre les seues parts. Aquesta metodologia existeix tot i que la seua aplicació no està generalitzada.

La metodologia composicional consisteix en tres passos bàsics: 1) la transformació de les composicions a altres tipus de coordenades; 2) l'ajust d'un model de regressió a les dades

transformades mitjançant l'aplicació de les ferramentes estadístiques clàssiques i la predicció de valors pel model en aquestes coordenades; 3) la transformació inversa que permet recuperar les estimacions en forma de composició.

Cal assenyalar que la metodologia exposada necessita que totes les parts de la composició siguen estrictament positives. Açò fa necessari detectar els possibles valors nuls a la sèrie de dades original i actuar en conseqüència. El tipus de zeros que podem trobar a la sèrie de dades del JMP són zeros que anomenem arrodonits i la solució passa per reemplaçar-los per un valor petit que permeta aplicar la metodologia sense alterar ni la sèrie ni el model ajustat a aquesta. La tècnica de substitució ha de garantir conservar la condició de suma constant. A aquesta tesina hem utilitzat l'estratègia de substitució multiplicativa.

L'elecció del valor de substitució adequat per a cada país no és senzilla i la generalització d'un valor a escala regional o per a un dels tipus de dades que tenim (aigua o sanejament en context rural o urbà) precisa d'un anàlisi en profunditat que excedeix el propòsit d'aquesta tesina. A la tesina es presenta el problema i s'analitza la influència que té l'adopció de dos valors de substitució distints sobre dos països en particular, respecte a les dades d'accés a aigua i al sanejament,. En un dels casos els resultats obtinguts són equivalents i en l'altre molt distints.

Tal i com hem vist, la interpretació de la sensibilitat del model al valor de substitució escollit és immediata si analitzem l'ajust en coordenades transformades donat que s'aprecia com valors molt petits suposen una variació de la pendent de la recta d'ajust que es tradueix en grans variacions de curvatura del model en l'espai del Símplex.

Caldria aprofundir en l'anàlisi de sensibilitat dels resultats en funció del valor de substitució adoptat, per a cada país i per a cada sèrie de dades (aigua o sanejament i context rural o urbà). Aquest anàlisi per país hauria de determinar el valor de substitució més adequat. Obtinguts aquests valors caldria comparar-los per determinar l'existència d'un valor comú a escala geogràfica (regional, mundial) o en funció del tipus de dades (accés a l'aigua o al sanejament) i/o de la subpoblació (entorn rural o urbà) de la qual es tracte.

Una vegada resolta l'estratègia de tractament dels zeros arrodonits per a les sèries de dades, l'aplicació de l'anàlisi composicional a les dades del JMP és, a més de necessari, senzill, fàcilment sistematitzable, respectuós amb la natura de les dades i ens permet realitzar prediccions a escenaris futurs qualsevol que siga l'escenari temporal fixat.

11 AGRAÏMENTS

La realització d'aquesta tesina no haguera estat possible sense el suport de les institucions i persones següents:

- Joint Monitoring Program, que ens facilità la seua base de dades i ens donà el suport tècnic necessari per tal de poder-la interpretar correctament.
- Al Grup de Recerca en Cooperació i Desenvolupament Humà (GRECDH) de la UPC, i especialment a Agustí Pérez-Forguet i a Ricard Giné Garriga sense el suport dels quals aquesta tesina no existiria.
- A Maribel Ortego Martinez, tutora d'aquesta tesina i suport imprescindible per a la realització d'aquesta.

12 BIBLIOGRAFIA

Aitchison, J. (1982). The Statistical Analysis of Compositional Data. *Journal of the Royal Statistical Society. Series B. Methodological*, *44*(2), 139–177. doi:10.2307/2345821

Aitchison, J. (1984). Reducing the dimensionality of compositional data sets. *Journal of the International Association for Mathematical Geology*, *16*(6), 617–635. doi:10.1007/BF01029321

Aitchison, J., & Egozcue, J. J. (2005). Compositional Data Analysis: Where Are We and Where Should We Be Heading? *Mathematical Geology*, *37*(7), 829–850. doi:10.1007/s11004-005-7383-7

Anderson, D. R., & Burnham, K. P. (2002). *Model selection and multimodel inference. A practical Information-Theoretic Approach*. (Springer, Ed.) (2nd ed., Vol. 53). doi:10.1017/CBO9781107415324.004

Bartram, J., Brocklehurst, C., Fisher, M. B., Luyendijk, R., Hossain, R., Wardlaw, T., & Gordon, B. (2014). Global monitoring of water supply and sanitation: history, methods and future challenges. *International Journal of Environmental Research and Public Health*, *11*(8), 8137–65. doi:10.3390/ijerph110808137

Egozcue, J. J., & Pawlowsky-Glahn, V. (2011a). Análisis composicional de datos en Ciencias Geoambientales. *Boletin Geologico Y Minero*, *122*(4), 439–452.

Egozcue, J. J., & Pawlowsky-Glahn, V. (2011b). Basic Concepts and Procedures. *Compositional Data Analysis: Theory and Applications*, 12–28. doi:10.1002/9781119976462.ch2

Fuller, J. A., Goldstick, J., Bartram, J., & Eisenberg, J. N. S. (2016). Tracking progress towards global drinking water and sanitation targets: A within and among country analysis. *Science of the Total Environment*, *541*, 857–864. doi:10.1016/j.scitotenv.2015.09.130

Hamed, A. (2008). *Guía del Mundo 2008. El presente y sus razones*. (E. SM, IEPALA, & I. del T. Mundo, Eds.). Madrid.

Martín-Fernández, J. A., Barceló-Vidal, C., & Pawlowsky-Glahn, V. (2003). Dealing With Zeros and Missing Values in Compositional Data Sets Using Nonparametric Imputation. *Mathematical Geology*, *35*(3), 253–278. Retrieved from http://www.google.de/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0CD0QFjAB&url=http://www.researchgate.net/publication/226223129_Dealing_with_Zeros_and_Missing_Values_in_Compositional_Data_Sets_Using_Nonparametric_Imputation/file/9fcfd50c21

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2012). Model-based replacement of rounded zeros in compositional data: Classical and robust approaches. *Computational Statistics and Data Analysis*, *56*(9), 2688–2704. doi:10.1016/j.csda.2012.02.012

Montgomery, D. ., Peck, E. ., & Vining, G. G. (2015). *Introduction to linear regression analysis*. (Wiley&Sons, Ed.)*Introduction to linear regression analysis*. doi:10.3174/ajnr.A2184

Pawlowsky-Glahn, V., & Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. *Compositional Data Analysis in the Geosciences: From Theory to Practice*, *264*, 1–10. doi:10.1144/GSL.SP.2006.264.01.01

PNUD. (2006). Fin de la crisis de agua y saneamiento. In *Informe Mundial de Desarrollo Humano 2006, titulado "Más allá de la escasez: poder, pobreza y la crisis mundial",* (p. 50).

R Development Core Team, R. (2011). *R: A Language and Environment for Statistical Computing*. *R Foundation for Statistical Computing* (Vol. 1). doi:10.1007/978-3-540-74686-7

Shapiro, S. S., & Wilk, M. B. (1965). Biometrika Trust An Analysis of Variance Test for Normality (Complete Samples). *Source: Biometrika Biometrika Trust*, *52*(34), 591–611. doi:10.1093/biomet/52.3-4.591

WHO, & UNICEF. (2014). WHO/UNICEF JMP Task Force on Methods, 1–39. doi:10.1038/sj.embor.7401032

WHO/UNICEF. (n.d.). The JMP uses households surveys and censuses. Retrieved September 22, 2016, from http://www.wssinfo.org/definitions-methods/data-sources/

WHO/UNICEF. (2015). 2015 Update and MDG Assessment. Retrieved from http://www.who.int/water_sanitation_health/publications/jmp-2015-update/en/

Wolf, J., Bonjour, S., & Prüss-Ustün, A. (2013). An exploration of multilevel modeling for estimating access to drinking-water and sanitation. *Journal of Water and Health*, *11*(1), 64–77. doi:10.2166/wh.2012.107

Zou, K. H., Tuncali, K., & Silverman, S. G. (2003). Correlation and simple linear regression. *Radiology*, *227*(3), 617–622. doi:10.1148/radiol.2273011499

Saisana et al (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. Journal of the Royal Statistical Society Series a-Statistics in Society, 168, 307–323.

van den Boogaart, K. G., & Tolosana-Delgado, R. (2008). "compositions": A unified R package to analyze compositional data. Computers and Geosciences, 34(4), 320–338. doi:10.1016/j.cageo.2006.11.017